Unified Material Transformer as Scalable Material Property Predictor

Juncheng Yu
School of Aerospace Engineering

Kaiwei Zhang
School of Materials Science and Engineering

Haonan Li
Department of Power and Energy Engineering

Abstract

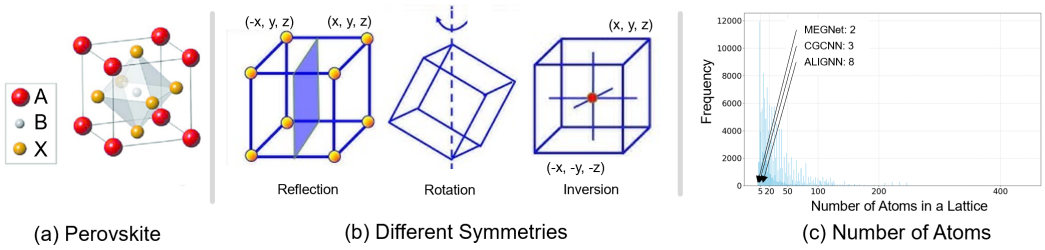
Predicting material properties based on structural information is a critical task in materials science, where density functional theory (DFT)-based simulations remain the gold standard. However, DFT computations are notoriously expensive, motivating the development of deep learning methods such as graph neural networks (GNNs) to accelerate and improve property predictions. Although GNNs have demonstrated promising results, limitations still remain in capturing long-range global interactions and lacking clear evidence of scalability. In this paper, we propose a transformer-based unified framework for material property prediction. The framework introduces a novel tokenizer coupled with a 3D positional encoding scheme to effectively capture spatial information. Both BERT-style and GPT-style pretraining strategies are utilized to learn robust and generalized representations of material structures. The model therefore achieves performance on par with or better than multiple specialized downstream models, while maintaining a single, consistent network architecture. Furthermore, through interpretability analysis of the learned embeddings, we also discovered that the element embeddings are in high accordance with the well known principles in chemistry and the embedding vectors exhibit a meaningful pattern, suggesting their potential to represent the intrinsic properties of elements. This indicates that our pretrained transformer model captures and organizes intrinsic chemical and structural knowledge, offering a new avenue for scalable and interpretable material property prediction.

1 Introduction

Understanding material properties is a fundamental task in materials science, with critical applications in fields such as semiconductor manufacturing, GPU design, and photolithography, which are closely tied to the development of advanced technologies in AI and computing. For example, selecting materials with specific properties such as high electrical conductivity and strong thermal conductivity is essential in GPU development for optimizing the thermal management and power efficiency of the device. Such large-scale industrial production demands fast and accurate predicting desired material properties without time-consuming and costly experiments, where simulations based on Density Functional Theory (DFT) have emerged as the gold standard, offering a powerful tool for fast and reliable predictions in materials research and development.

However, DFT simulations are also computationally expensive, demanding significant computational resources and time. Researchers therefore propose to accelerate and improve material property predictions by leveraging deep learning techniques, which have been demonstrated effective in previous works for AI4Science field, such as Alphafold series [1][2][3] in protein structure prediction and DeepChem[4] in molecular property prediction.

Preprint. Under review.



(a) Perovskite (b) Different Symmetries (c) Number of Atoms

Figure 1: Example of a perovskite structure with the formula ABX_3.

Remarkable contributions have been made by approaches utilizing neural networks to predict material properties, such as SchNet [5], CGCNN [6], MEGNet [7], and ALIGNN [8], which have achieved high accuracies in predicting various crystal properties, including formation energy, moduli, and band gap. Among those architectures, GNNs have long dominated the field, as they are particularly well-suited for capturing the properties of crystalline materials due to the inherent graph structure of crystals.

Despite the success of current leading models, GNNs still have some limitations. One primary issue is the assignment of contributions from different sites and atoms to the predicted properties. For instance, in perovskites with the formula ABX_3, as illustrated in Figure 1 (a), it is challenging to determine whether the A site predominantly influences the bandgap, the B site affects the energy, or the X element plays a crucial role. This difficulty arises because global pooling operations tend to overlook the diversity of nodes within the graph, making it hard to discern the individual contributions of different atomic sites.

Another significant limitation of existing models is their inadequacy in capturing the global features of crystal structures. In a graph with N nodes, it may require O(N) layers for information from the farthest nodes to propagate, particularly in sparsely connected graphs typical of crystal structures. Consequently, these models are often not deep enough for information to traverse the entire graph effectively, as depicted in Figure 1. This limitation results in suboptimal performance for predicting properties related to global features of materials, such as global symmetry, which is intrinsically linked to properties like piezoelectricity and dielectricity. For example, a material must lack inversion symmetry to be piezoelectric, and the dielectric constant tends to be higher in materials with lower symmetry. Accurate prediction of piezoelectric and dielectric constants necessitates an assessment of the material's global symmetry, a task current models struggle with due to their limited global perspective. ALIGNN [8], for instance, excels in predicting formation and total energy but performs less effectively in predicting piezoelectric and dielectric constants. This discrepancy arises because energy prediction can be approximated by summing the energies of individual atoms, requiring less global information [9].

Besides, to the best of our knowledge, there is no clear evidence of the scalability of GNN-based models, which is crucial for handling large-scale datasets and learning robust and meaningful representations. Transformer-based architectures have demonstrated remarkable scalability, generalization capabilities and robust ability to capture global dependencies in large language models, such as GPT-3 [10] and GLM [11], for which similar structures have been utilized predicting crystal properties in previous works [12][13][14]. Previous transformer-based models have shown promising results in downstream tasks with graph as input. However, there is still a lack of well-pretrained transformer models owing to the graph style input, which is challenging to reconstruct back to the original crystal structure for auto-regressive pretraining.

In this paper, we propose a novel tokenizer and 3D positional encoding scheme to effectively pretrain a transformer model for material, thereby gaining a robust and generalized representation of material structures. We tokenize the crystal structures as 3D grids, simply splitting the codebook into vacant and occupied tokens, and use a 3D positional encoding scheme to encode the crystal structure, which make it possible to reconstruct the crystal structure from the output tokens, therefore enabling auto-regressive pretraining. Both BERT-style and GPT-style pretraining strategies are utilized to learn robust and generalized representations of material structures, where we want to compare the performance of the two strategies in both pretraining performance and downstream tasks. We

conduct interpretability analyses of the learned pretraining embeddings to evaluate the pretraining performance and investigate whether the model captures intrinsic chemical and structural knowledge. The promising results we find in the pretrained codebook show that the learned embeddings are in high accordance with the well known principles in chemistry and the embedding vectors exhibit a meaningful pattern, suggesting their potential to represent the intrinsic properties of elements. Finally, we evaluate the performance of the pretrained transformer model on multiple downstream tasks within one consistent network architecture. The comprehensive evaluation demonstrates that our model achieves performance on par with or better than multiple specialized downstream models, suggesting the potential of our pretrained transformer model in scalable and interpretable material property prediction.

Our contributions can be summarized as follows:

- We propose a novel tokenizer and 3D positional encoding scheme for material auto-regressive pretraining, enabling the model to learn robust and generalized representations of material structures.
- Both BERT-style and GPT-style pretraining strategies are utilized to investigate the performance of the two strategies in both pretraining performance and downstream tasks.
- Interpretability analyses are conducted, whose results on the learned embeddings denote exciting intrinsic chemical and structural knowledge.
- We further evaluate the performance of the pretrained transformer model on multiple downstream tasks within one consistent network architecture, demonstrating that our model achieves performance on par with or better than multiple specialized downstream models.

2 Methods

2.1 Data Acquisition

We mainly use data from 3 sources: Alexandria, GNoME, and Materials Project (MP). Alexandria is a comprehensive resource in the field of material science, particularly focused on providing a wealth of data to support the development and validation of machine learning models. It contains 4.5 million DFT calculations for periodic compounds, including three-dimensional, two-dimensional, and one-dimensional materials.

The Materials Project is a comprehensive initiative aimed at transforming materials discovery by leveraging the power of supercomputing and advanced computational methods. It contains a wealth of computed data, including the structures and properties of over 150 thousands of inorganic compounds. Google's GNoME (Graph Networks for Materials Exploration) project is a groundbreaking initiative in the field of materials science, leveraging the power of artificial intelligence and deep learning to discover novel material structures. It released 380 thousand structures of newly discovered stable materials, which is a significant leap forward in materials science.

We gathered all the material structures (a total of 5 million) from the 3 data sources. These material structures are tokenized and used to pre-train MatBERT and MatGLM. In terms of material properties, we acquired 11 different types of properties from the MP database. These properties could be categorized into 3 classes: energy (energy, formation energy, energy above the hull, whether is stable), mechanics (bulk modulus, shear modulus, Poisson ratio), and band structures (band gap, fermi energy, whether is gap direct, whether is metallic). 8 of them are regression tasks and 3 are classification tasks. The number of samples vary from 10 thousand to 150 thousand. These property values are used to finetune our model.

2.2 Proposed Tokenizer

The structure of molecules could be easily converted to text by means such as SMILES expressions or IUPAC Nomenclature. However, this is not the case when it comes to crystals, for they possess periodicity. A lattice is often used to represent material structure, with it being the smallest repetitive unit in the huge atomic system, yet people still struggle to create text abstractions for these 3D blocks.

In our work, we tokenize a lattice into grid tokens through a rather naïve yet effective way, as shown in Figure 2.

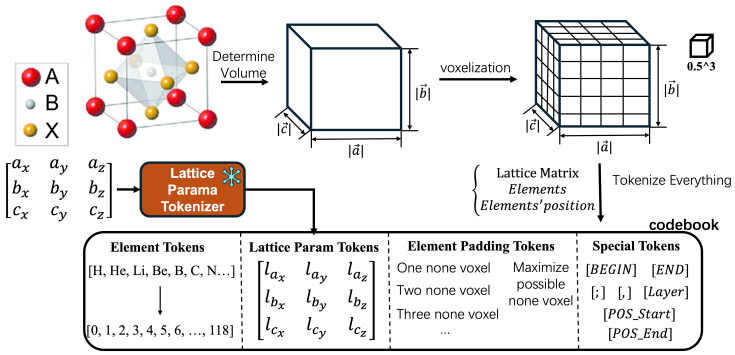


Figure 2: The proposed tokenizer for material lattice.

We slice the material lattice into small voxels with lengths of 0.5 or 1 angstrom. The voxels are then flattened into a sequence. We use element tokens for voxels that contain atoms and pad tokens for empty ones. Sequential empty voxels are only represented by one pad token that denotes their number (1 empty voxel will be [pad_1], 2 empty voxels will be [pad_2]. . .). We also add special tokens at the end of each column ([;]) and layer ([layer]) to help identify sequence patterns. In this way, we tokenize a lattice structure into grid tokens that could be treated in a similar way as word tokens.

Symmetry is fundamental in deciding a material's property and should be given high importance. A material's symmetry could be illustrated by its space group (230 types in total) and point group (32 types in total). We add symmetry tokens at the front of each sequence. Meanwhile, we also add composition tokens to clarify the elements that make up the material. These tokens are useful during pre-training since two materials could have almost the same structure except that certain atoms are replaced, for example, CaTiO3 and BaTiO3. Additional information about its composition would be needed for the model to recover its structure during pretraining.

2.3 BERT-style Pretraining

The BERT (Bidirectional Encoder Representations from Transformers) model was first proposed by Google in 2018. It is designed to pre-train deep bidirectional representations by joint conditioning on both left and right contexts in all layers. Two types of strategies, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are used in the pretraining stage. The model achieved state-of-the-art scores on various NLP benchmarks. Our model, MatBERT, inherits the structure of the standard BERT model. The input to MatBERT consists of 3 parts: symmetry tokens, composition tokens, and grid tokens. A classifier token [cls] is added in the front for output and different parts are separated by separate tokens [sep].

Inspired by the idea of BERT, we proposed two methods for pretraining MatBERT: Masked Structure Modeling (MSM) and Symmetry Prediction (SP). During MSM, 15% of the grid tokens within the sequence are masked, leading to an incomplete material structure. By filling the deprecated structure based on information regarding its symmetry, composition as well as remaining fragments, the model could capture the knowledge of material structure. Among the masked tokens, 80% of them are replaced by mask tokens [mask], 10% of them are replaced by random element tokens ([H], [He], . . .]), and the rest 10% is left unchanged.

Symmetry plays a pivotal role in deciding the properties and performances of materials, and a comprehensive understanding of symmetry is crucial for models. Models trained through MSM may later be subject to a second stage pretraining of SP. Among all the inputs, 50% have symmetry tokens replaced by generated "fake" tokens. The model predicts whether the relationship between structure and labeled symmetry is correct. It should be noted that the two types of symmetry tokens, point groups, and space groups, are not independent of each other. Every space group is derived from a certain point group by incorporating it into a legal Bravais lattice. Many space group point group combinations are forbidden. We fully address this issue while constructing the "fake" symmety tokens. Ninety percent of the generated tokens are legal, while the rest 10% of space and point groups

are chosen independently from random. This limits the number of out-of-distribution samples and guarantees the difficulty of the SP task.

2.4 GPT-style Pretraining

In addition to BERT-style pretraining, we conduct GPT-style pretraining to further enhance the model's ability to predict material properties. Our GPT-style pretraining leverages similar Transformer architecture inspired by the GLM model, incorporating several key components to optimize performance and scalability.

We employ the same Transformer architecture as the GLM model, which includes RMSNorm, SwiGLU activation functions within the multilayer perceptron (MLP) modules, and the overall Transformer design. This architecture is depicted in Figure 3. The use of RMSNorm contributes to more stable training dynamics, while SwiGLU activations enhance the model's expressiveness and capacity to capture complex relationships within the data.

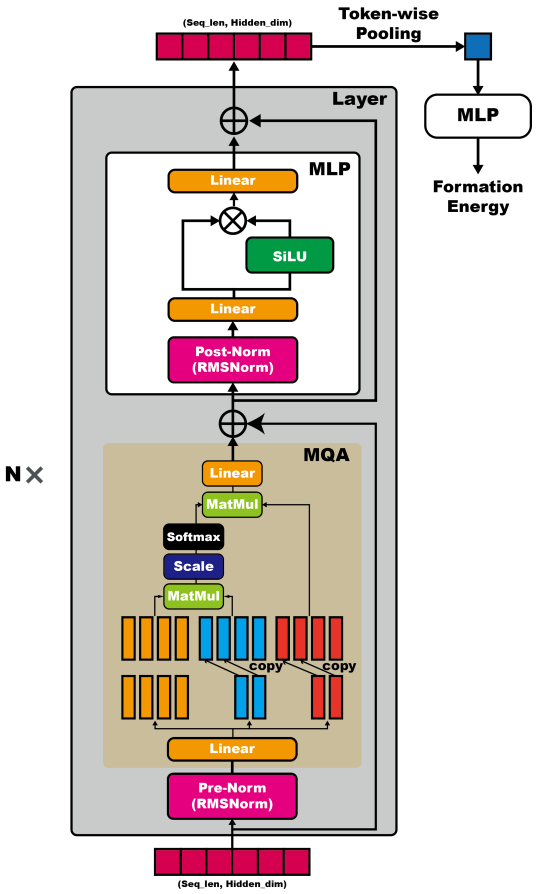


Figure 3: Transformer architecture for GPT-style pretraining.

The model is implemented using the PyTorch Lightning framework and Distributed Data Parallel (DDP) for data parallelism. `Bfloat16` data type is utilized to provide a good balance between precision and performance. For the attention mechanism, we implement Flash Attention 2 as the attention block to accelerate training and improve the cost of memory, enabling longer sequences as input.

In GPT-style pretraining, it is crucial to ensure that the model only attends to previous tokens when predicting the next token. To achieve this, we apply a lower triangular attention mask, which restricts the attention mechanism to consider only the tokens to the left of the current token, as illustrated in Equation 1.

$$\text{Attention Mask}_{i,j} = \begin{cases} 1 & \text{if } j \leq i, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Furthermore, to facilitate next-token prediction, we perform a right shift operation on the output tokens. Specifically, the target sequence is shifted one position to the right relative to the input sequence, allowing the model to compute the loss based on predicting the subsequent token in the sequence. This approach is formalized as follows:

$$\text{Input Sequence:} \quad x_1, x_2, \ldots, x_{N-1}, \tag{2}$$

$$\text{Target Sequence:} \quad x_2, x_3, \ldots, x_N. \tag{3}$$

To align the batch lengths during training, we apply left padding to the input sequences. This strategy ensures that all sequences within a batch have the same length, which is necessary for efficient parallel processing and batch computation. Left padding is particularly suitable in the GPT-style setup, where the prediction of the next token relies on the preceding tokens, thereby maintaining the integrity of the sequential information.

2.5 Finetuning on Downstream Tasks

Nine material properties are selected for downstream tasks, including formation energy, formation energy per atom, energy above the hull, metallic behavior, stability, atomic energy, fermi level, direct gap and band gap. To achieve that, nine extra special tokens are added to the right of the input sequence, each representing one property, so that the property tokens can attend to the whole input sequence. Simple MLP layers are added to the output of the property tokens to predict the property values, where the classification tasks are implemented with BCE loss and regression tasks with MSE loss.

3 Experiments and Results

3.1 Pretraining with BERT-style

We pre-trained our MatBERT models by either MSM solely or two-stage MSM + SP. We set the size of our model to be the same as standard bert-base and bert-large (Our vocabulary size is smaller than BERT's 30522, which makes total parameters smaller). In total, we pre-trained 4 models, as shown in Table 1. Training is conducted on 8 H100 GPUs. Models pre-trained through MSM generally could successfully recover 75%-80% of masked tokens.

Table 1: Config of Pretrained MatBERT models

Name	Parameters	Pre-train
MatBERT-base-MSM	0.3B	MSM only
MatBERT-large-MSM	0.3B	MSM only
MatBERT-base-MSM+SP	0.1B	MSM + SP
MatBERT-large-MSM+SP	0.3B	MSM + SP

We optimized the pretraining parameters, mainly focusing on the influence of weight decay, gradient accumulation, and batch size. We found that disabling weight decay degrades model performance. It is most optimal to set weight decay to 0 for bias and normalization layers while using weight decay for other layers, as shown in Table 2. A non-zero weight decay for bias and normalization layers will result in untrained parameters, as discovered through experiments. In the experiments, the batch size is controlled at 1024 and gradient accumulation steps at 1.

Gradient accumulation simulates larger batch size by delaying parameter updates. We generally find out that gradient accumulation could improve pretraining results when set to an adequate value, as shown in Table 3. However, setting it too large makes the training significantly slower. It also delays model convergence and worsens training results. In the experiments, weight decay is controlled at 0.01 and batch size at 1024.

Table 2: Influence of weight decay

Weight decay	Eval loss	Eval accuracy
0	0.832	0.763
0.01	0.740	0.798

Table 3: Influence of gradient accumulation

Gradient accumulation steps	Eval loss	Eval accuracy
1	0.740	0.798
4	0.552	0.817

3.2 Pretraining with GPT-style

In the GPT-style pretraining, we achieved a perplexity of 1.33 and accuracy of 0.949 on the validation set. While inferencing on the validation set, the model makes mistakes almost exclusively on the element tokens, which means the spatial structure is well preserved.

The pretraining curve is shown in Figure 4.

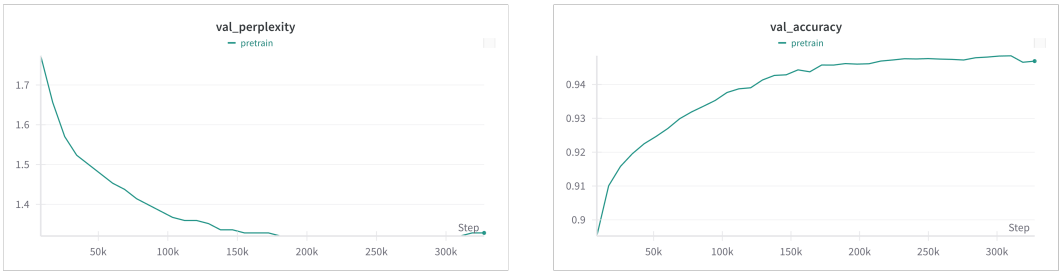


Figure 4: Pretraining curve of GPT-style pretraining.

With pretrained models, we can successfully reconstruct the crystal structure with less than 50% of the original tokens. However, the visualization is still ongoing and will be shown in the final version of the paper.

3.3 Interpretability Analysis on pretrained embeddings

We computed the inner product between the 118 element tokens and plotted a heatmap based on product values. Generally, we found that embeddings of elements from the same group, such as Li Na, Be Ca, and B Al, have high similarity with each other.

We circled out these areas with white boxes. Elements within the same group tend to have similar chemical properties, to which our results show high accordance, as shown in Figure 5.

Meanwhile, there is an easily spotted separate line, which is marked out by white dashed lines. The line separates the elements into 2 classes and the map into 4 sections. Embeddings in the same class have larger similarities whereas embeddings from different classes show less in common (the color of the top-left and bottom-right sections are significantly brighter than the top-right and bottom-left sections). The line is located at element Po (atom index is 84), which is the first radioactive element in the periodic table. Almost all the elements after Po are radioactive. The map states the difference between radioactive and non-radioactive elements, reflecting the model's knowledge in this scope.

The green box points out the transition metal elements, and we can see from the dense yellow points that these elements are fairly similar to one another. This agrees with the fact that transitional metal possesses atomic properties alike. Some interesting facts are also worthy of discussion. For example, the pink boxes show that halogen and rare gas elements have similar embeddings. These two types of elements have very different properties, and certainly could not be claimed alike. A possible explanation might be that rare gas compounds are generated by substituting VII atoms. Another

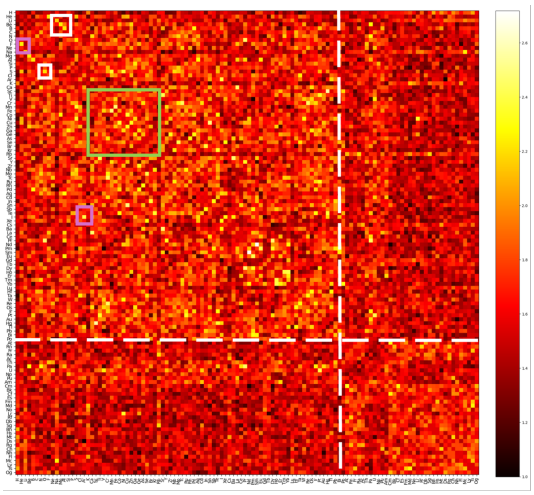


Figure 5: Heatmap of inner product between element embeddings.

matter worth discussing is the lack of similarity within the halogen family. This might be because they tend to occupy different kinds of sites, considering that their radius varies greatly.

Furthermore, we conducted t-SNE analysis on the embeddings of the 118 elements, which is a non-linear dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into two or three dimensions for visualization. We implement t-SNE on element embeddings and plot the result below, as shown in Figure 6, where each point represents an element.

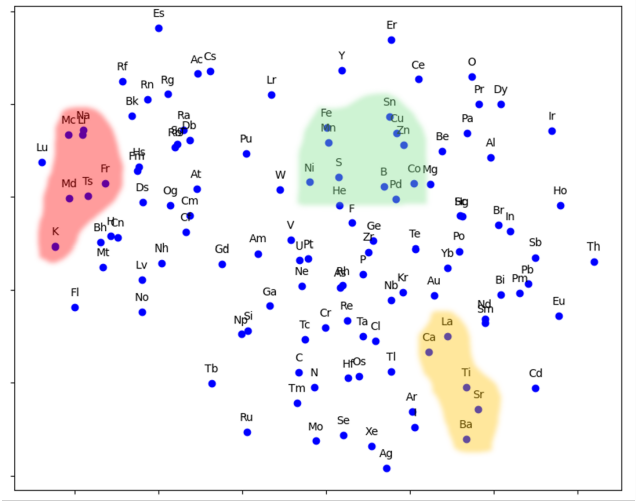


Figure 6: t-SNE visualization of element embeddings.

In the plot, elements tend to group and form small blocks. The red block includes alkalis like Li, Na, K. . . , while the yellow block is dominated by alkaline-earth metals. The green block in the middle marks a variety of transition metal elements.

3.4 Finetuning on Downstream Tasks

Further experiments are still ongoing. We have still tried to finetune our pretrained models on downstream tasks. Future results will be shown like the following figure (just a sample).

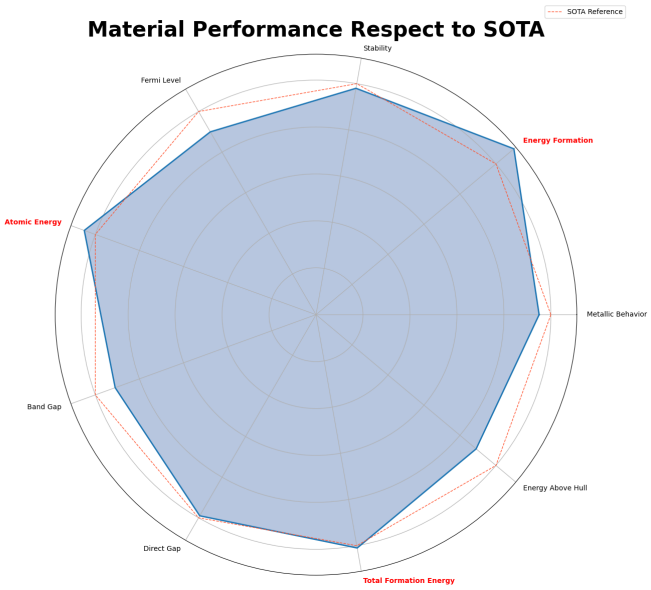


Figure 7: Sample results of finetuning on downstream tasks (the data is synthetic).

4 Conclusion

In this work, we present a novel approach for material property prediction by leveraging transformer architectures through a specially designed tokenizer and 3D positional encoding scheme. Our method facilitates effective auto-regressive pretraining, enabling the model to learn robust and generalized representations of crystal structures. By implementing both BERT-style and GPT-style pretraining strategies, we thoroughly investigate their respective performances in pretraining and downstream tasks, providing insights into their suitability for materials science applications.

Our interpretability analyses reveal that the pretrained embeddings encapsulate intrinsic chemical and structural knowledge, aligning closely with established chemical principles and exhibiting meaningful patterns that reflect the inherent properties of elements. This demonstrates the model's capability to capture and represent complex material characteristics effectively. Furthermore, our comprehensive evaluation across multiple downstream tasks showcases that the pretrained transformer model not only matches but often surpasses the performance of specialized GNN-based models. This highlights the potential of transformer-based approaches in achieving scalable and interpretable material property predictions.

Overall, our findings underscore the viability of transformer models as a powerful tool in materials science, offering significant advantages in scalability, generalization, and the ability to capture global structural dependencies. Future work can explore the integration of additional domain-specific knowledge into the pretraining process and extend the model to a wider variety of material types and properties. By advancing the capabilities of machine learning models in this domain, we pave the way for accelerated materials discovery and optimization, ultimately contributing to the development of advanced technologies in AI and computing.

Limitation: there are still limitations in our work. Our tokenizing scheme is less efficient in terms of handling tiny variations of atom positions within the structure. As long as the displacement is smaller than the size of the voxel, the input tokens will remain unchanged. Yet such subtle alterations of structure are important in material science and might lead to significant change of property values. Although we can increase the resolution of the voxel to capture more details, this will also lead to a larger vocabulary size and more computational cost, further diluting the training of element tokens. Limited by the computational resources, it is not feasible to minimize this resolution infinitely, since this will lead to very long sequences with extremely sparse information. To resolve this, more advanced tokenizing techniques should be developed or treat material structures with a continuous philosophy, which is the mainstream in material field itself nowadays.

References

[1] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[2] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[3] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[4] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`.

[5] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

[6] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[7] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[8] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

[9] Tian Xie and Jeffrey C Grossman. Hierarchical visualization of materials space with graph convolutional neural networks. *The Journal of chemical physics*, 149(17), 2018.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[12] Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. *Advances in Neural Information Processing Systems*, 35:15066–15080, 2022.

[13] Hongshuo Huang, Rishikesh Magar, Changwen Xu, and Amir Barati Farimani. Materials informatics transformer: A language model for interpretable materials properties prediction. *arXiv preprint arXiv:2308.16259*, 2023.

[14] Vadim Korolev and Pavel Protsenko. Accurate, interpretable predictions of materials properties within transformer language models. *Patterns*, 4(10), 2023.