Boost Your Crystal Model with Denoising Pre-training

Shuaike Shen [*1] **Ke Liu** [*1] **Muzhi Zhu** [1] **Hao Chen** [1]

Abstract

Crystals play a vital role in a wide range of materials, influencing both cutting-edge technologies and everyday applications. Recently, deep learning approaches for crystal property prediction have shown exceptional performance, driving significant progress in material discovery. However, supervised approaches can only be trained on labeled data and the number of data points varies for different properties. Making full use of unlabeled data remains an ongoing challenge. To address this issue, we propose an unsupervised Denoising Pre-training Framework (DPF) for crystal structure. DPF trains a model to reconstruct the original crystal structure from recover the masked atom types, perturbed atom positions, and perturbed crystal lattices. Through the pre-training, models learn the intrinsic features of crystal structures and capture the key features influencing crystal properties. We pre-train models on 380,743 unlabeled crystal structures and fine-tune them on downstream property prediction benchmarks. Extensive experiments demonstrate the effectiveness of our denoising pre-training framework.

1. Introduction

Crystals play a pivotal role in a diverse range of materials, including cutting-edge materials like superconductors and everyday application like solar materials (Kittel et al., 1996). Precise prediction of crystal properties is paramount for material discovery and advancement of society. Physics theory guarantees that the structure of a crystal profoundly influences its properties, which sheds light on the modeling of crystal structures with geometric deep learning (LeSar, 2013). Benefiting from publicly available data from physical experiments and *in silico* simulations, crystal models based on geometric deep learning have been vigorously developed

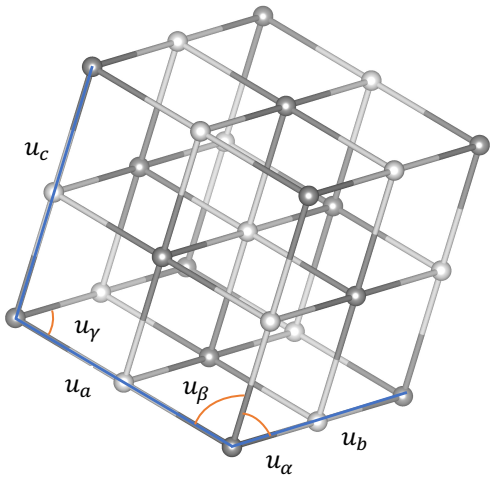

Figure 1. Illustration of a crystal unit cell structure. The circles represent the atoms and their respective locations. The parallelepiped indicates the Bravais lattice. The lattice vectors, depicted as red lines, determine the orientation and periodicity of the lattice.

(Choudhary & DeCost, 2021; Xie & Grossman, 2018; Liu et al., 2022; Yan et al., 2022; Chen et al., 2019a; Liu et al., 2024).

However, the labeled data for crystal property prediction are notably scarce and the number of data points varies with different properties. For example, **JARVIS** (Choudhary et al., 2020) only consists of 55,714 labeled crystals. For properties, like *Shear Moduli*, there are only 4,664 entries. Comparing to the 14,197,122 annotated data in the ImageNet (Russakovsky et al., 2015) dataset in the field of computer vision, such amount of data is quite not enough to train a deep learning model. Therefore, making full use of unlabeled data is essential for crystal property prediction. Moreover, these target properties lack comparability, suggesting a lack of specific relationships between them. Consequently, deep learning models must discern distinct patterns for each target property, posing a challenge to comprehensively modeling the crystal structure.

Recently, deep learning models have achieved significant breakthroughs in material discovery and design (Merchant et al., 2023; Cui et al., 2024). (Merchant et al., 2023), employing deep learning tools, has discovered 2.2 million potential crystal structures. These massive crystal structures

[*]Equal contribution [1]Zhejiang University, China. Correspondence to: Hao Chen <haochen.cad@zju.edu.cn>.

5th AI for Science workshop at the 41^{st} *International Conference on Machine Learning (ICML)*, 2024.

discovered by deep learning tools adhere to the chemical and physical principles governing crystal structures, which make it possible to pre-train a foundation model to learn the general pattern inside the structures. We propose a **D**enoising **P**re-training **F**ramework (**DPF**) to leverage these unlabeled data for crystal property prediction. Specifically, we perturb the lattice, atom types, and atom positions of crystal representations. Then perturbed crystal structures are fed into the denoising neural network to reconstruct the original structure. Subsequently, we fine-tune the pre-trained model for downstream crystal property prediction tasks.

The main contributions of our work can be summarized as follows:

- We propose a novel denoising pre-training framework (DPF), which extracts general patterns and key features of crystal structures. Our DPF acts as a foundation for various downstream crystal property prediction tasks.
- We provide an in-depth analysis of the relationship between the pre-training performance and different perturbing techniques, including the lattice, atom positions, and atom types.
- Extensive experimental results across widely recognized crystal property prediction benchmarks demonstrate the effectiveness of our DPF.

2. Related Works

2.1. Crystal Material Property Prediction

The prediction task for crystal material properties was initially based on their chemical formulas (Villars & Phillips, 1988; Stanev et al., 2018; Jha et al., 2018; 2019; Wang et al., 2021) . (Villars & Phillips, 1988) and (Stanev et al., 2018) utilize machine learning techniques such as support vector machine (SVM) and random forests to predict crystal properties by analyzing the statistics of their compositions. (Jha et al., 2018), (Jha et al., 2019), and (Wang et al., 2021) treat chemical formulas as sentences and apply sequence models to them. As both composition and structure significantly influence crystal properties, recent methods have shifted focus towards modeling the three-dimensional structure of crystals (Xie & Grossman, 2018; Liu et al., 2022; Chen et al., 2019a; Choudhary & DeCost, 2021). In their CGCNN model, Xie & Grossman (2018) introduced a multiedge graph, where atoms serve as nodes and edges are drawn between atom pairs based on manually defined distances. Following CGCNN, MEGet proposed a global node to capture environmental information (Chen et al., 2019b). Subsequently, ALIGNN incorporated atom angle information into MEGet (Choudhary & DeCost, 2021). Additionally, Matformer (Yan et al., 2022) encoded periodic patterns by considering geometric distances between atoms with identical type in neighboring cells.

2.2. Denoising Pre-training

To leverage unlabeled data effectively, previous works proposed several denoising pre-training techniques for computer vision (CV), natural language processing (NLP), and molecules (Han et al., 2021). In the field of NLP, techniques such as masked language models (MLM) and token replacement detection are frequently employed during the pre-training stage (Radford et al., 2018; Devlin et al., 2018). Auto-encoders (Bank et al., 2023), a staple in computer vision, encompass various forms, such as the Masked Auto-Encoder (He et al., 2022), adept at capturing highly compressed image information. Contrastive learning is commonly utilized to effectively leverage unlabeled data for discerning differences between positive and negative samples (Caron et al., 2021; Radford et al., 2021; He et al., 2020). In our framework, models are pre-trained on extensive unlabeled datasets with denoising tasks and subsequently fine-tuned on datasets labeled with crystal properties. We notice that in a contemporaneous work, Song et al. (2024) also tried to pre-train models with a diffusion process, which requires fractional coordinates as input for their models. Different from them, our framework is more general and can be applied to any model without changing its architecture. Besides, we pre-train the models with a much larger dataset, which leads to much better performance.

2.3. Datasets

The benchmark for crystal property prediction is well established. There are two standard benchmark datasets, JARVIS and Materials Project (Choudhary et al., 2020; Chen et al., 2019b), to evaluate the performance of our model. The two datasets are widely used in various works (Yan et al., 2022; Chen et al., 2019a; Xie & Grossman, 2018), which include labeled data for the common properties like formation energy, band gap, bulk moduli, and so on. Merchant et al. (2023) have discovered 2.2 million crystal structures using a deep generative tool. After filtering the repetitive and physically or chemically irrational structures, we utilize 380,743 structures to pre-train our model.

3. Preliminary and Notations

The crystal structure $C = \{B, L\}$ can be effectively characterized by a basis B and a Bravais lattice L (Kittel et al., 1996). As depicted in Fig. 1, the Bravais lattice $L \in \mathbb{R}^6$ is represented by a parallelepiped, defined by six lattice constants, *i.e.*, the lengths of its three edges and the angles between them $L = \{u_a, u_b, u_c, u_\alpha, u_\beta, u_\gamma\}$. The basis $B = \{A, P\}$ comprises atom types A and their respective positions P. By repeating the basis in the direction

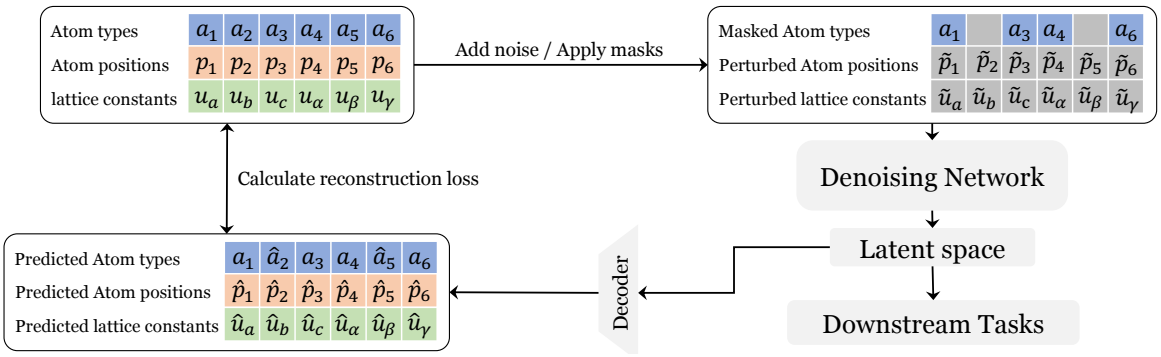

Figure 2. The pipeline of our denoising pre-training framework. At the pre-training stage, we apply mask on crystal atom types, add noise on atom positions and lattice constants to acquire perturbed representations. Then we put the perturbed crystal representations into denoising network to reconstruct the corresponding original representations. Subsequently, the learned structure latent space can be fine-tuned for downstream tasks like crystal property prediction.

of the Bravais lattice edges, the entire crystal structure is generated. Fig. 1 shows the unit cell, which is single periodic unit in the crystal internal structure. The atom types $A = \{a_1, a_2, \cdots, a_n\}$ consist of the types of all atoms in the crystal, where $a_i \in \mathcal{C}^{95}$. $\mathcal{C}^{95}$ indicates 95 types of atoms. The atom positions $P = \{p_1, p_2, \cdots, p_n\}$ consist of the positions of all atoms in the basis, where $p_i \in \mathbb{R}^3$ indicates the 3D position of an atom.

The crystal property prediction problem involves predicting the property of a given crystal structure $C = \{A, P, L\}$ by learning a function f to predict its corresponding property.

4. Denoising Pre-training Framework

In this section, we introduce our denoising pre-training framework (DPF) in detail. DPF consists of mask atom type modeling, atom position perturbing, and lattice constant perturbation, as shown in Fig. 2.

Mask language modeling. For mask language modeling, we randomly select γ of atoms in a crystal and assign a type *unknown* to them following Liu et al. (2022) as follows:

$$\tilde{A} = \{\mathbb{1}(a_1, \epsilon_1), \mathbb{1}(a_2, \epsilon_2), \cdots, \mathbb{1}(a_n, \epsilon_n)\}, \quad (1)$$

where $\mathbb{1}$ is the indicator, where ϵ_i is 1 if the atom is selected and the $\mathbb{1}(a_i, 1)$ is assigned the type unknown. Finally, the model is trained to predict the ground truth types of them with the position of them only.

Atom position denoising. For each atom in a crystal, we randomly sample a noise $\eta \in \mathbb{R}^3 \sim \mathcal{N}(0, 1)$ and apply it to the atom positions as follows:

$$\tilde{P} = \{p_1 + \alpha \cdot \eta_1, p_2 + \alpha \cdot \eta_2, \cdots, p_n + \alpha \cdot \eta_n\}, \quad (2)$$

where $\alpha \in \mathbb{R}$ indicates the scalar of noise. Finally, the model is trained to predict the true atom positions using the perturbed atom positions.

Lattice parameter denoising. Lattice parameters are essential components of crystals and indicates their periodicity. We also perturb the lattice constants with a randomly sampled Gaussian noise $\sigma \in \mathbb{R}^6 \sim \mathcal{N}(0, 1)$ as follows:

$$\tilde{L} = L + \beta \cdot \sigma \quad (3)$$

where β is the lattice parameter noise scalar.

With the three denoising self-supervised tasks, the model is pre-trained to capture the intrinsic features of crystal structure representations.

5. Experiments

To verify the effectiveness and generalization performance of our pre-training framework, we fine-tune the model on two commonly used benchmarks, *i.e.*, JARVIS and Materials Project. Besides, we tested the impact of different mask ratios on the model's prediction of crystal properties.

5.1. Experimental Setup

Datasets. We pre-train our model on 380,743 crystal structure data filtered from the recent work (Merchant et al., 2023), excluding structures that are duplicates of downstream datasets or do not have physical or chemical significance. We mainly conduct experiments on two standard benchmarks, **JARVIS** (Choudhary et al., 2020) and **Materials Project** (Chen et al., 2019b). The JARVIS dataset contains 55,722 materials with critical crystal properties for functional material design, including bandgaps, formation energies, energy above hull, total energy and so on. The Materials Project dataset aggregates several key crystal property datasets, including formation energy, band gap, bulk moduli, and shear moduli. Among them, 69,239 crystals

Method	Formation Energy↓ eV/atom	BandGap(OPT)↓ eV	Total Energy↓ eV/atom	Ehull↓ eV	Bandgap(MBJ)↓ eV
CFID*	0.14	0.30	0.24	0.22	0.53
CGCNN*	0.063	0.20	0.078	0.17	0.41
SchNet*	0.045	0.19	0.047	0.14	0.43
MEGNet*	0.047	0.145	0.058	0.084	0.34
GATGNN*	0.047	0.17	0.056	0.12	0.51
ALIGNN*	0.0331	0.142	0.037	0.076	0.31
Matformer*	0.0325	0.137	0.035	0.064	**0.30**
DPF(γ = 30%)	**0.029**	**0.118**	0.0289	<u>0.036</u>	0.311
DPF(γ = 50%)	<u>0.031</u>	<u>0.122</u>	**0.0286**	**0.035**	0.315
DPF(γ = 70%)	**0.029**	0.123	<u>0.0288</u>	<u>0.036</u>	0.316

Table 1. The experimental results in terms of MAE on JARVIS dataset. * denotes the results are taken from the referred papers. The best results are shown in bold and the sub-optimal results are underlined.

are labeled with properties of formation energy and band gap, while only 5,451 crystal structures are labeled with the properties of bulk moduli and shear moduli. More details are in Appendix. A.

Compared approaches. We compare our DPF with previous crystal property prediction models, including **CFID** (Choudhary et al., 2018), **CGCNN**, (Xie & Grossman, 2018), **SchNet** (Schütt et al., 2017), **MEGNet** (Chen et al., 2019b), **GATGNN** (Louis et al., 2020), **ALIGNN** (Choudhary & DeCost, 2021), **Matformer** (Yan et al., 2022). Matformer explicitly consider the lattice constant as the periodicity of the crystal. ALIGNN is a typical crystal models used for simulating the interactions between atoms in crystals.

Implementation details. All our experiments are conducted on computing clusters with GPUs of NVIDIA® GeForce® RTX 4090 24GB and CPUs of AMD® EYPC® 7542 CPU @ 2.90GHz. We pre-train our model for 50 epochs with optimizer *AdamW*, batch size 256, learning rate 0.001, weight decay 10^{-5}, and one cycle scheduler. Subsequently, we fine-tune our model on downstream crystal property prediction tasks for 500 epochs, batch size 32 and all other hyperparameters are exactly the same as the pre-training stage.

Evaluation metrics. Following previous works (Choudhary & DeCost, 2021; Xie & Grossman, 2018; Chen et al., 2019a; Yan et al., 2022; Liu et al., 2024), we employ the Mean Absolute Error (**MAE**) to evaluate the accuracy of crystal property prediction.

$$\text{MAE}(C, f) = \frac{1}{m}\sum_{i=1}^{m}|f(C_i) - y_i|$$

5.2. Experimental results

5.2.1. Different mask ratios on atom types

The quantitative results on the Materials Project benchmark dataset are shown in Table. 2. Our DPF models with different mask ratios of atom types show significant improvements over baseline models on three out of four sub-tasks of the Materials Project benchmark. It is worth noting that the improvement of previous models on the *Bulk Moduli* task has consistently been very limited since it is particularly challenging for models to learn the key factors affecting the *Bulk Moduli* property from limited data. By pre-training on a large number of unlabeled crystal structures, our model has acquired the capability to capture crystal structure features and exhibits improved robustness. Specifically, our method outperforms the previous model by **2.8%** on the Bulk Moduli task.

The experimental results on the JARVIS benchmark dataset are shown in the Table. 1. Our models outperform the previous models on 4 out of 5 tasks, particularly achieving significant improvements on these four tasks. Specifically, **10.77%** on Formation Energy, **13.87%** on BandGap(OPT), **18.29%** on Total Energy and **45.31%** on Ehull.

Furthermore, by predicting the masked atom types, the pre-trained model can extract general features of the crystal representation, making it more robust and less sensitive to noise. We show the fine-tuning process of the pre-trained models in Fig. 3 and Fig. 4. We can see that by pre-training with masked atom types, our models have smoother training curves with less fluctuation and an earlier convergence point.

5.2.2. Different noise scale to lattice constants

The test results of the models pre-trained on different noise scale to lattice constants are shown in Table. 3. Our model

Method	Formation Energy↓ eV/atom	Band Gap↓ eV	Bulk Moduli↓ log(GPa)	Shear Moduli↓ log(GPa)
CGCNN*	0.031	0.292	0.047	0.077
SchNet*	0.033	0.345	0.066	0.099
MEGNet*	0.030	0.0307	0.060	0.099
GATGNN*	0.033	0.280	0.045	0.075
ALIGNN*	0.022	0.218	0.051	0.078
Matformer*	0.021	0.211	0.043	**0.073**
DPF(γ = 30%)	0.0201	<u>0.206</u>	0.0443	0.0738
DPF(γ = 50%)	**0.0196**	0.210	**0.0418**	<u>0.0734</u>
DPF(γ = 70%)	<u>0.0200</u>	**0.203**	<u>0.0424</u>	0.0756

Table 2. The experimental results in terms of MAE on Materials Project dataset. * denotes the results are taken from the referred papers. The best results are shown in bold and the sub-optimal results are underlined.

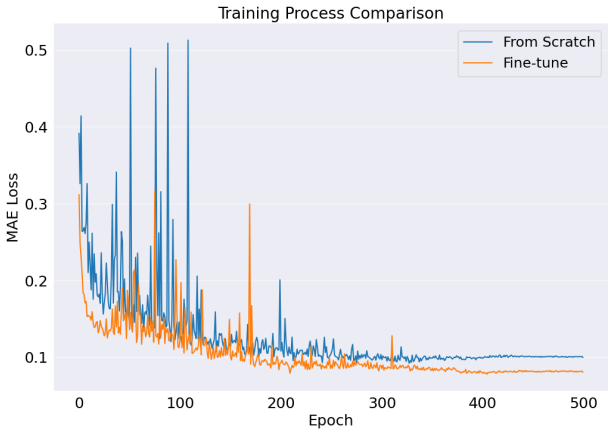

Figure 3. The comparison between from scratch training and fine-tuning from pre-trained model on *Band Gap* task of **Materials Project** benchmark.

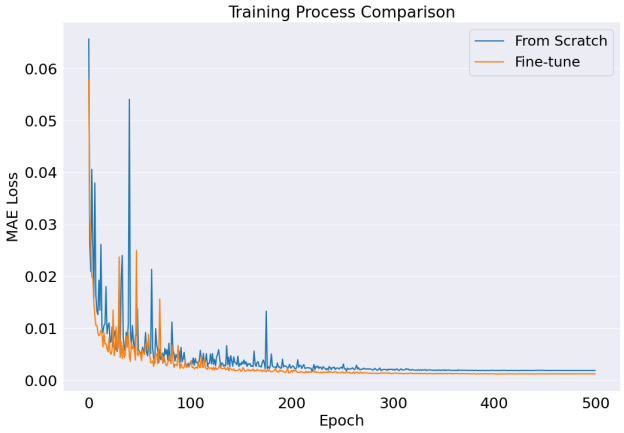

Figure 4. The comparison between from scratch training and fine-tuning from pre-trained model on *Total Energy* task of **JARVIS** benchmark.

outperforms the baseline in 6 out of 9 sub-tasks across two benchmarks, achieving a notable **16.05%** improvement on the BandGap(OPT) task. However, overall, the performance boost from pre-training with lattice constant noise is less significant and more unstable compared to pre-training with masked atom types. Furthermore, pre-training the model with lattice constant noise lead to varying degrees of performance decline compared to the baseline, which is quite unusual.

5.2.3. Different noise scale to atom positions

We also conduct pre-training by adding noise to the atom positions. The experimental results are presented in Table. 4. Our models outperform the baseline on 8 out of 9 sub-tasks. It is worth noting that this pre-training strategy achieves a **4.2%** improvement on the Bulk Moduli task, on which the model's performance is limited by the amount of training data. Moreover, our model pre-trained with $\alpha = 10\%$ performs better on *Shear Moduli* sub-task which other pre-training strategies cannot accomplish. However, this training strategy encounters the same issue as pre-training with noise on lattice constants, resulting in decreased model performance. For example, DPF(50% position) pre-training on atom positions with a 50% noise scale led to performance drops in the *Shear Moduli*, *Bulk Moduli*, and *Bandgap(MBJ)* sub-tasks compared to the baseline.

On other sub-tasks, is performance declined declines compared to when it is pre-trained with masked atom types.

5.2.4. Discussion on the results

Through observing and analyzing the fine-tuning results of different pre-training strategies, we find that perturbing crystal structure features (*i.e.*, lattice constants and atom positions) yield poorer performance compared to pre-

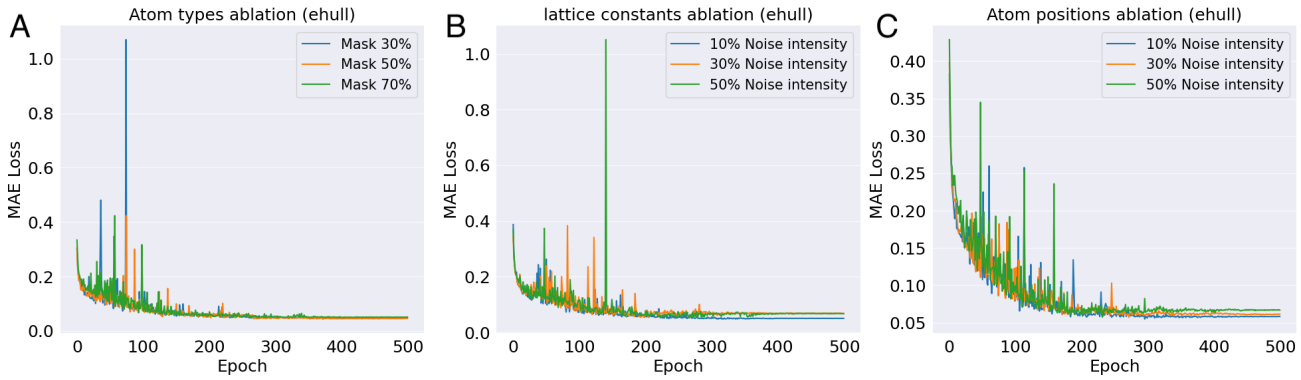

Figure 5. The fine-tuning process of the pre-trained models with different perturbation strategies for crystal structures on *ehull* task. **A)** denotes the fine-tuning process of different mask ratios on atom types. **B)** and **C)** are the training process of different noise scale to lattice constants and atom positions.

training with masked atom types. Additionally, the model pre-trained with masked atom types exhibits more stable performance, with the best results being more consistently achieved across various sub-tasks. We believe that for a crystal structure representation, reconstructing masked atom types based on lattice constants and atom positions is more straightforward. The model can infer unknown atom types by learning atomic radii and inter-atomic forces. Moreover, the correlation between crystal property prediction and atom type reconstruction tasks is stronger than that of reconstructing crystal structures, making the extracted structural features more useful.

However, it's worth noting that reconstructing perturbed structures has been proven effective in molecules and has been incorporated into various models. For example, adding translation and rotation noise to each amino acid in a protein and applying rotation noise to amino acid side chains (Jin et al., 2023; Zhang et al., 2024), helps train denoising networks. We conclude that the fundamental reason for this asymmetry lies in the differences in constructing graph representations. For molecules, when constructing graph representations, the edges of the graph are formally established based on physical connections, such as chemical bonds like C-C bonds or C-H bonds in proteins. However, for inorganic crystal structures, graph representations are built using the KNN (K-nearest neighbors) (Fix, 1985) algorithm, where edges are based on distances between atoms. Thus, when we perturb the lattice constants or atom positions, some edges in the graph may vanish while new edges are constructed, disrupting the overall graph representation. This makes it challenging to reconstruct the structure before perturbation, and features learned from the post-perturbation graph may no longer be applicable to downstream tasks such as crystal property prediction. As depicted in Fig .5A, when applying different mask ratios to atom types for pre-training, there is little difference in the loss curves during fine-tuning.

JARVIS	Formation Energy↓ eV/atom	BandGap(OPT)↓ eV	Total Energy↓ eV/atom	Ehull↓ eV	Bandgap(MBJ)↓ eV
Matformer	0.0325	0.137	0.035	0.064	**0.30**
DPF(β = 10%)	**0.029**	0.118	0.0293	**0.036**	0.329
DPF(β = 30%)	0.030	0.119	**0.0291**	0.037	0.329
DPF(β = 50%)	0.030	**0.115**	0.0298	0.037	0.326
Materials Project	Formation Energy↓ eV/atom	Band Gap↓ eV	Bulk Moduli↓ log(GPa)	Shear Moduli↓ log(GPa)	
Matformer	0.021	0.211	**0.043**	**0.073**	
DPF(β = 10%)	0.0197	0.212	0.0443	0.0744	
DPF(β = 30%)	0.0198	0.207	0.0456	0.0739	
DPF(β = 50%)	**0.0196**	**0.206**	0.0440	0.0745	

Table 3. The experimental results in terms of MAE on JARVIS dataset and Materials Project dataset of different perturbed scale to lattice constants.

JARVIS	Formation Energy↓ eV/atom	BandGap(OPT)↓ eV	Total Energy↓ eV/atom	Ehull↓ eV	Bandgap(MBJ)↓ eV
Matformer	0.0325	0.137	0.035	0.064	**0.30**
DPF(α = 10%)	**0.030**	0.122	**0.0292**	**0.038**	0.310
DPF(α = 30%)	0.031	**0.121**	0.0296	0.039	0.329
DPF(α = 50%)	0.031	0.126	0.0297	0.038	0.329
Materials Project	Formation Energy↓ eV/atom	Band Gap↓ eV	Bulk Moduli↓ log(GPa)	Shear Moduli↓ log(GPa)	
Matformer	0.021	0.211	0.043	0.073	
DPF(α = 10%)	0.0207	**0.210**	0.0432	**0.0722**	
DPF(α = 30%)	0.0203	0.211	**0.0411**	0.0755	
DPF(α = 50%)	**0.0201**	0.217	0.0443	0.0758	

Table 4. The experimental results in terms of MAE on JARVIS dataset and Materials Project dataset of different perturbed scale to atom positions.

However, when adding noise to lattice constants and atom positions, the fluctuation of the training loss curve increases with the scale of the noise. Particularly, the green curve in Fig 5C exhibits significant fluctuations, indicating that the crystal features extracted during pre-training are no longer suitable for downstream tasks. To further elucidate the impact of perturbations on downstream tasks regarding crystal structures, we introduced varying scales of noise to all crystal representations. The comparison of fine-tuning process is shown in Fig. 6. We can see that more severe perturbations to crystal structure, *i.e.*, lattice constants and atom positions, result in more pronounced fluctuations during the training process.

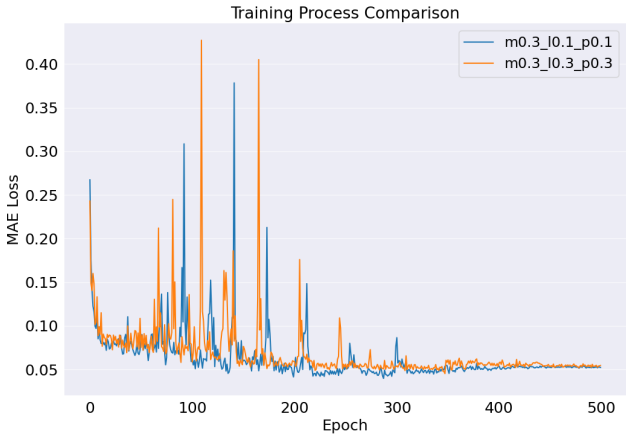

Figure 6. The comparison between the fine-tuning process of Mask 30% atom types, adding 10% noise scale to lattice constants, adding 10% noise scale to atom positions and fine-tuning process of Mask 30% atom type, add 30% noise scale to lattice constants, add 30% noise scale to atom positions on *Bulk Moduli* task.

6. Conclusion

In this work, we propose a denoising pre-training framework (DPF) to reconstruct the perturbed crystal structure representations, which can make full use of crystal data without property annotations. By reconstructing the crystal structure representations, the network can extract the general pattern and features of the crystal. Then we fine-tune the pre-trained model on the crystal property prediction tasks, and experimental results show that our DPF architecture outperforms the baseline models consistently. The denoising network in our proposed pre-training framework can be any equivariant featurizer, making our framework easily adaptable to any model. Additionally, further experiments show that pre-training by masking atom types performs better on downstream tasks compared to pre-training by perturbing crystal spatial structure parameters. We believe this is because perturbing the spatial structure parameters disrupts the graph representation.

Meanwhile, our work has some limitations. We did not explore downstream tasks involving crystal structure generation, where pre-training by perturbing spatial structure parameters might be more effective and could enhance the diversity of generated crystals.

References

Bank, D., Koenigstein, N., and Giryes, R. Autoencoders. Machine learning for data science handbook: data mining and knowledge discovery handbook, pp. 353–374, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 31(9): 3564–3572, 2019a.

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 31(9): 3564–3572, 2019b.

Choudhary, K. and DeCost, B. Atomistic line graph neural network for improved materials property predictions. npj Computational Materials, 7(1):185, 2021.

Choudhary, K., DeCost, B., and Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. Physical review materials, 2(8):083801, 2018.

Choudhary, K., Garrity, K. F., Reid, A. C., DeCost, B., Biacchi, A. J., Hight Walker, A. R., Trautt, Z., Hattrick-Simpers, J., Kusne, A. G., Centrone, A., et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. npj computational materials, 6(1):173, 2020.

Cui, T., Tang, C., Su, M., Zhang, S., Li, Y., Bai, L., Dong, Y., Gong, X., and Ouyang, W. Geometry-enhanced pretraining on interatomic potentials. Nature Machine Intelligence, pp. 1–9, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Fix, E. Discriminatory analysis: nonparametric discrimination, consistency properties, volume 1. USAF school of Aviation Medicine, 1985.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., et al. Pre-trained models: Past, present and future. arXiv e-prints, pp. arXiv–2106, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009, 2022.

Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., and Agrawal, A. Elemnet: Deep learning the chemistry of materials from only elemental composition. Scientific reports, 8(1):17593, 2018.

Jha, D., Ward, L., Yang, Z., Wolverton, C., Foster, I., Liao, W.-k., Choudhary, A., and Agrawal, A. Irnet: A general purpose deep residual regression framework for materials discovery. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2385–2393, 2019.

Jin, W., Chen, X., Vetticaden, A., Sarzikova, S., Raychowdhury, R., Uhler, C., and Hacohen, N. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. bioRxiv, pp. 2023–12, 2023.

Kittel, C., McEuen, P., and McEuen, P. Introduction to solid state physics, volume 8. Wiley New York, 1996.

LeSar, R. Introduction to computational materials science: fundamentals to applications. Cambridge University Press, 2013.

Liu, K., Yang, K., Zhang, J., and Xu, R. S2snet: A pretrained neural network for superconductivity discovery. In Raedt, L. D. (ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pp. 5101–5107. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/708. URL https://doi.org/10.24963/ijcai.2022/708. AI for Good.

Liu, K., Yang, K., and Gao, S. A periodicity aware transformer for crystal property prediction. Neural Computing and Applications, pp. 1–12, 2024.

Louis, S.-Y., Zhao, Y., Nasiri, A., Wang, X., Song, Y., Liu, F., and Hu, J. Graph convolutional neural networks with global attention for improved materials property prediction. Physical Chemistry Chemical Physics, 22(32): 18141–18148, 2020.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. Nature, 624(7990):80–85, 2023.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115: 211–252, 2015.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems, 30, 2017.

Song, Z., Meng, Z., and King, I. A diffusion-based pre-training framework for crystal property prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 8993–9001, 2024.

Stanev, V., Oses, C., Kusne, A. G., Rodriguez, E., Paglione, J., Curtarolo, S., and Takeuchi, I. Machine learning modeling of superconducting critical temperature. npj Computational Materials, 4(1):1–14, 2018.

Villars, P. and Phillips, J. C. Quantum structural diagrams and high-t c superconductivity. Physical Review B, 37 (4):2345, 1988.

Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., and Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. Npj Computational Materials, 7(1):77, 2021.

Xie, T. and Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters, 120 (14):145301, 2018.

Yan, K., Liu, Y., Lin, Y., and Ji, S. Periodic graph transformers for crystal material property prediction. Advances in Neural Information Processing Systems, 35:15066–15080, 2022.

Zhang, Z., Xu, M., Lozano, A. C., Chenthamarakshan, V., Das, P., and Tang, J. Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. Advances in Neural Information Processing Systems, 36, 2024.

A. Dataset Statistics

A.1. Pretraining Dataset

The statistical information from the filtered pre-training dataset is shown in the Table. A.2.1 below.

Task	Metric	$\|\mathbf{C}\|$	$\|\mathbf{A}\|$	$\|\mathbf{T}\|$	Volume	Density
Pre-training Dataset	Max	380740	40	6	9291.69	24.10
	Min		2	2	25.84	0.18
	Mean		21.46	4.10	436.96	8.34
	Var		90.06	0.46	62283.63	7.38

Table 5. Statistics of the filtered pre-training datasets. Max, Min, Mean, and Var are the maximum value, the minimum value, the average, and the variance of the data respectively. $|\mathbf{A}|$, $|\mathbf{T}|$, and $|\mathbf{C}|$ indicate the number of atoms per crystal, atom types per crystal, and number of crystal structures.

A.2. Benchmark Dataset

A.2.1. Materials Project Dataset

The statistical information from the Materials Project is shown in the Table. A.2.1 below.

Task	Metric	$\|\mathbf{C}\|$	$\|\mathbf{A}\|$	$\|\mathbf{T}\|$	Volume	properties
Formation Energy	Max	55712	140	7	8904.04	4.99
	Min		1	1	5.66	-4.42
	Mean		10.09	2.92	178.97	-0.83
	Var		82.09	0.5	37490.46	1.17
Bandgap (OPT)	Max	55712	140	7	8904.04	9.64
	Min		1	1	5.66	0.00
	Mean		10.09	2.92	178.97	0.69
	Var		82.09	0.5	37490.46	1.99
Total Energy	Max	55712	140	7	8904.04	3.39
	Min		1	1	5.66	-10.5
	Mean		10.09	2.92	178.97	-3.21
	Var		82.09	0.5	37490.46	4.54
Ehull	Max	55712	140	7	8904.04	8.33
	Min		1	1	5.66	0.00
	Mean		10.09	2.92	178.97	1.78
	Var		82.09	0.5	37490.46	1.94
Bandgap (MBJ)	Max	55712	140	7	8904.04	27.5
	Min		1	1	5.66	0.00
	Mean		10.09	2.92	178.97	1.50
	Var		82.09	0.5	37490.46	5.38

Table 6. Statistics of the Materials Project datasets. Max, Min, Mean, and Var are the maximum value, the minimum value, the average, and the variance of the data respectively. $|\mathbf{A}|$, $|\mathbf{T}|$, and $|\mathbf{C}|$ indicate the number of atoms per crystal, atom types per crystal, and number of crystal structures.

A.2.2. JARVIS dataset

The statistical information from the JARVIS is shown in the Table. A.2.2 below.

Task	Metric	$\lvert\mathbf{C}\rvert$	$\lvert\mathbf{A}\rvert$	$\lvert\mathbf{T}\rvert$	Volume	properties
Formation Energy	Max	69239	296	8	6901.60	4.39
	Min		1	1	5.60	-4.52
	Mean		29.91	3.32	469.95	-1.65
Band Gap	Max	69239	296	8	6901.60	17.89
	Min		1	1	5.60	0.00
	Mean		29.91	3.32	469.95	1.35
	Var		810.94	0.83	241066.28	2.63
Bulk Moduli	Max	5450	152	6	2396.02	2.64
	Min		1	1	5.60	0.48
	Mean		9.50	2.57	158.73	1.94
	Var		83.79	0.34	22515.70	0.13
Shear Moduli	Max	5449	152	6	2396.02	2.72
	Min		1	1	5.60	0.30
	Mean		9.50	2.57	158.73	1.62
	Var		83.81	0.34	22519.79	0.14

Table 7. Statistics of the JARVIS datasets. Max, Min, Mean, and Var are the maximum value, the minimum value, the average, and the variance of the data respectively. $\lvert\mathbf{A}\rvert$, $\lvert\mathbf{T}\rvert$, and $\lvert\mathbf{C}\rvert$ indicate the number of atoms per crystal, atom types per crystal, and number of crystal structures.

B. Training Details

All our experiments are conducted on computing clusters with GPUs of NVIDIA® GeForce® RTX 4090 24GB and CPUs of AMD® EYPC® 7542 CPU @ 2.90GHz. We pre-train our model on 4 RTX 4090 GPUs for 50 epochs with optimizer *AdamW*, batch size 256, learning rate 0.001, weight decay 10^{-5}, and one cycle scheduler. Subsequently, we fine-tune our model on downstream crystal property prediction tasks on one RTX 4090 GPU for 500 epochs, batch size 32 and all other hyperparameters are exactly the same as the pre-training stage.

C. Implementation Details

The detailed model architecture is shown in Fig 7. The denoising network can be replaced with any feature extraction network, and the downstream tasks can include crystal property prediction, classification, and crystal generation. This flexibility gives our architecture strong transferability. In our paper, we use Matformer as the denoising network, which has achieved excellent test results on downstream benchmark tasks.

D. Additional Results

The contemporaneous work, CrysDiff (Song et al., 2024) also use the denoising process to pre-train its model and subsequently fine-tune the pre-trained model. The comparison between CrysDiff and out models are shown in Table. 8. As both methods involve pre-training followed by fine-tuning, the comparison between our DPF model and CrysDiff is fair. As shown in the Table. 8, when perturbing the crystal spatial structure, i.e., lattice constants and atom positions), our DPF methods and CrysDiff exhibit varying performance across different sub-tasks. When pre-training with masked atom types, our DPF method outperforms the CrysDiff model in four out of five tasks. Additionally, our approach does not require the complex and tedious conversion of fractional coordinates. By using the original model as a feature learner, our method is simpler in architecture and can be easily transferred to other models.

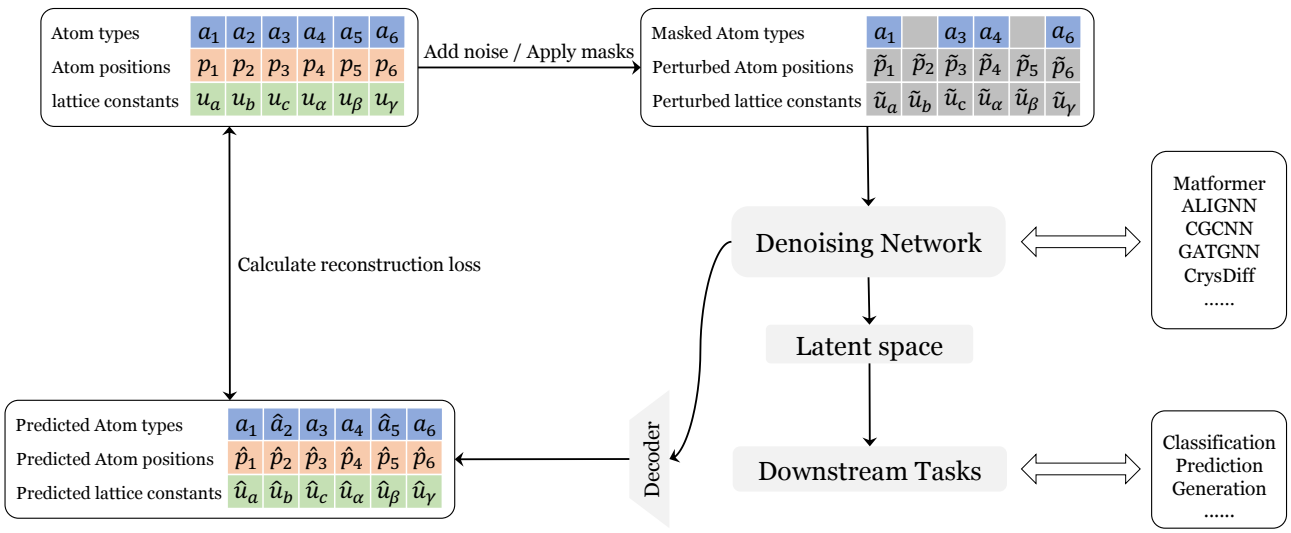

Figure 7. The pipeline of our denoising pre-training framework. The denoising network component can be replaced with any feature extraction network. The downstream tasks can include crystal property prediction, classification, and crystal generation.

Method	Formation Energy↓ eV/atom	BandGap(OPT)↓ eV	Total Energy↓ eV/atom	Ehull↓ eV	Bandgap(MBJ)↓ eV
CrysGNN*	0.056	0.183	0.069	0.130	0.371
CrysDiff*	**0.029**	0.131	0.034	0.062	**0.287**
DPF(γ = 30%)	0.029	0.118	0.0289	0.036	0.311
DPF(γ = 50%)	0.031	0.122	**0.0286**	**0.035**	0.315
DPF(γ = 70%)	**0.029**	0.123	0.0288	0.036	0.316
DPF(β = 10%)	0.029	0.118	0.0293	0.036	0.329
DPF(β = 30%)	0.030	0.119	0.0291	0.037	0.329
DPF(β = 50%)	0.030	**0.115**	0.0298	0.037	0.326
DPF(α = 10%)	0.030	0.122	0.0292	0.038	0.310
DPF(α = 30%)	0.031	0.121	0.0296	0.039	0.329
DPF(α = 50%)	0.031	0.126	0.0297	0.038	0.329

Table 8. The experimental results in terms of MAE on JARVIS dataset. * denotes the results are taken from the referred papers. The best results are shown in bold and the sub-optimal results are underlined.