A Geometric Foundation Model for Crystalline Material Discovery

Shengchao Liu[*1], **Divin Yan**[*2], **Weitao Du**[3], **Zhuoxinran Li**[4], **Zhiling Zheng**[1], **Omar Yaghi**[1], **Christian Borgs**[1], **Hongyu Guo**[5], **Anima Anandkumar**[2], **Jennifer Chayes**[1]
[1]University of California, Berkeley [2]California Institute of Technology [3]DAMO Academy
[4]University of Toronto [5]National Research Council Canada [*]Equal Contribution

Abstract

The use of artificial intelligence in crystalline material discovery is gaining significant attention from both the machine learning and chemistry communities. In this work, we present NeuralCrystal, a foundation model specifically designed to push the boundaries of material discovery by combining cutting-edge geometric modeling and large-scale pretraining techniques. The model ensures rotational and translational equivariance by using a vector frame basis, while projecting the coordinate system into the Fourier domain to capture the periodic symmetries and long-range interactions characteristic of crystalline materials. For geometric pretraining, we adopt an equivariant denoising approach by constructing dual views of crystalline structures from the Cambridge Structural Database. NeuralCrystal was rigorously tested on eight MatBench property prediction tasks, outperforming six, and demonstrating its strong potential to significantly accelerate the discovery of new materials. The codes are available at this GitHub repo.

1 Introduction

A foundation model is essential for crystalline material discovery. It enables a unified framework capable of generalizing across multiple tasks, eliminating the need to develop task-specific models and improving data efficiency. In material discovery, tasks such as predicting material energy and forces [3, 16, 32], material structure prediction [2, 29, 31], material generation [19, 28], and material optimization [17, 21, 43, 44] often share underlying principles. A foundation model can capture these commonalities, significantly improving efficiency and performance by learning a generalized geometric representation of crystalline materials. This shared knowledge allows the model to excel in multiple tasks without the need for task-specific tuning, accelerating the material discovery process.

The most critical part of such a foundation model lies in the geometric representation of crystalline materials. The mainstream geometric representation research line has been focusing on exploring geometric modeling on small molecules and proteins, treating them as a set of point clouds in the 3D Euclidean space [23, 36, 42]. Such representation function needs to be equivariant to rotation and translation, *i.e.*, SE(3)-equivariance. However, when modeling crystalline materials, one crucial property that crystal representations must satisfy is periodicity invariance, a feature that has been relatively under-explored in the research community. [1]

Specifically, crystalline structures exhibit long-range order and periodicity in their atomic or molecular arrangements, which should be effectively captured to fully represent their intrinsic geometric properties. This periodic arrangement is represented by a *lattice*, which is an array of points showing

[1]We leave the detailed discussion of SE(3)-equivariance to these works [4, 23, 36, 42], and in this work, we will be mainly focusing on the modeling of periodicity symmetry.

AI for Accelerated Materials Design Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

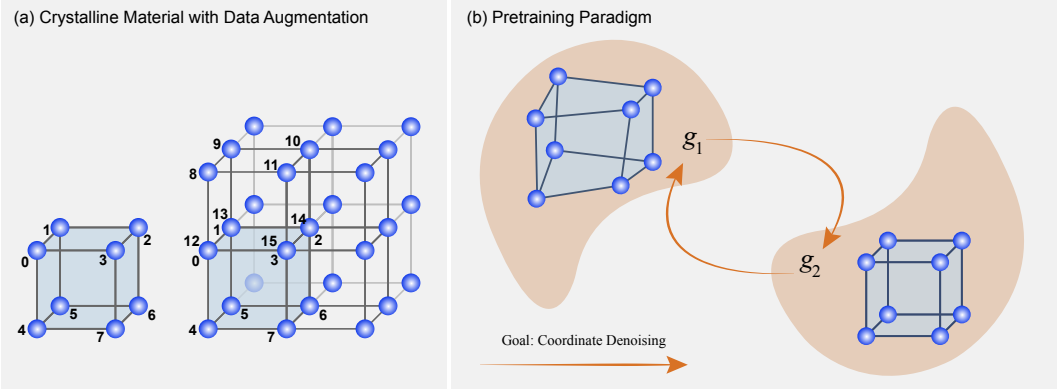

Figure 1: Illustration of NeuralCrystal. (a) Depicts the data structure for crystalline material (left) and data augmentation (right). (b) Illustrates the GeoSSL-PFM algorithm.

the potential positions of atoms or molecules. The periodicity of the crystal structure is defined by the repeating pattern of the lattice and the specific arrangement of atoms or molecules within the lattice. In essence, the lattice describes the overall periodic pattern of the crystal structure, while periodicity refers to the regular, repeating arrangement of atoms or molecules within that lattice pattern. The crystal is the physical manifestation of this periodic lattice arrangement of atoms or molecules. We illustrate this in Figure 1.

Existing works have mainly relied on data augmentation techniques to capture the periodic structures in crystals [23, 39]. The core idea involves augmenting a central unit cell by shifting it along the three lattice axes. Message passing [13] is then applied between atom pairs, with at least one atom positioned in the central cell. We illustrate this in Figure 1. However, relying on data augmentation to address periodic symmetry is still an approximation, and more efficient and more accurate methods are needed to handle this complexity effectively.

Meanwhile, there is a rich amount of high-quality material datasets with wet lab verifications, such as the Cambridge Structural Database (CSD) [14]. The supervised labels (*e.g.*, energies, and band gaps) are lacking on these datasets, yet the rich structure information is useful to help train a foundation model. This can be achieved using the technique called unsupervised pretraining or self-supervised learning [30, 33]. Existing pretraining methods are either supervised on materials [3, 34, 40] or unsupervised but focused on general molecules rather than crystalline materials [9, 24–26]. As a result, a tailored approach is needed to unlock the full potential of structural data for material discovery.

Our contributions. To this end, we propose NeuralCrystal, a geometric foundation model tailored for material discovery. Our approach presents two key innovations. (1) We first introduce FourierFrameNet, an SE(3)-equivariant and periodicity symmetric geometric representation for crystalline materials. This transformer-based model effectively captures long-range interactions by projecting pairwise atomic distances into the Fourier domain. (2) We then propose a novel geometric self-supervised pretraining method, GeoSSL-PFM, which embraces coordinate flow matching. The main idea is to add small perturbations to the geometric structures of crystalline materials, and then train a flow-matching model to maximize the mutual information between the original and perturbed geometries. To verify the effectiveness of NeuralCrystal, we empirically test NeuralCrystal on eight material property prediction tasks from MatBench, and the quantitative results show that NeuralCrystal reaches state-of-the-art performance on six of them. This reveals the potential of NeuralCrystal for achieving more versatile tasks in the future.

2 FourierFrameNet with SE(3)-Equivariance and Periodicity Symmetry

FourierFrameNet leverages vector frame for intra unitcell modeling (Section 2.1) and captures the periodicity at the inter unitcell level through the Fourier domain (Section 2.2).

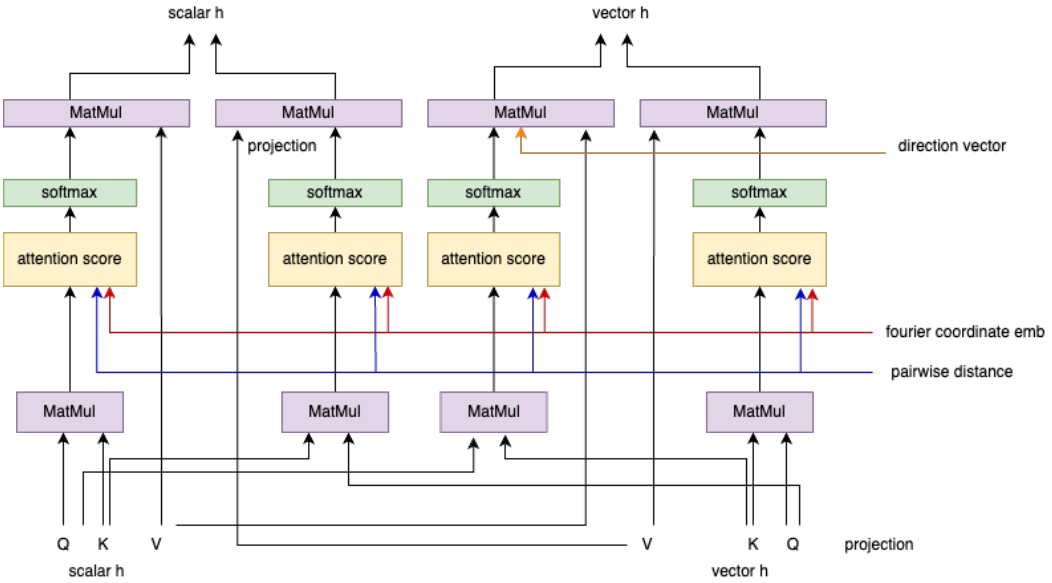

Figure 2: Illustration of periodic graph using FourierFrameNet Transformer.

2.1 Intra-cell Model: FrameNet

Vector frame basis construction. As summarized in recent works [23], utilizing the vector frame basis is an efficient and effective way for geometric modeling on molecules. Specifically in a molecular system like unit cell in the crystal structure, we extract all the atom pairs (i, j) within cutoff threshold c. For each atom i, we find the mass center of all its neighborhood atoms within the cutoff c, *i.e.*, $\boldsymbol{x}_k = \text{mean}(\boldsymbol{x}_j), \forall \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq c$. Then the atom-level vector frame is built as follows:

$$\mathcal{F} = \Big[\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2\Big] = \Big[\frac{\boldsymbol{x}_i - \boldsymbol{x}_k}{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|}, \frac{\boldsymbol{x}_i \times \boldsymbol{x}_k}{\|\boldsymbol{x}_i \times \boldsymbol{x}_k\|}, \frac{\boldsymbol{x}_i - \boldsymbol{x}_k}{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|} \times \frac{\boldsymbol{x}_i \times \boldsymbol{x}_k}{\|\boldsymbol{x}_i \times \boldsymbol{x}_k\|}\Big], \tag{1}$$

where $\times$ is the cross-product, and $\mathcal{F}_{i=0}^2 \in \mathbb{R}^{3\times 1}$ are the three bases in the vector frame. Notice that to guarantee the model is equivariant to translation, we also remove the mass center for each unit cell.

Notice that such a frame construction Equation (1) is constructed in each unit cell, *i.e.*, intra-cell level. If we expand this to the inter-cell, then there will be a numerical issue, since $\boldsymbol{x}_k$ can be very close to $\boldsymbol{x}_i$.

SE(3)-Equivariant Architecture. Then based on such an atom-level vector frame, we can build an SE(3)-equivariant geometric model. More concretely, we have two node-level representations on scalars and vectors, and they can be used to predict type-0 and type-1 quantities, respectively.

Assume the latent dimension is d, and the scalar representation is initialized as $\boldsymbol{h}_0^s = \text{one-hot}(\boldsymbol{x}) \in \mathbb{R}^{1\times d}$ and the vector representation is initialized with 0-value, as $\boldsymbol{h}_0^v = 0^{1\times 3\times d}$. We also adopt the radial basis function [6] multiplied by the cosine cutoff function to represent the atom pairwise distance, $\boldsymbol{h}^{\text{dist}} \in \mathbb{R}^{1\times d}$. The radial basis function is defined as:

$$f(r)^{\text{RBF}} = [e^{-\gamma(r-s_0)^2}, e^{-\gamma(r-s_1)^2}, ...], \tag{2}$$

where $s_0, s_1, ...$ are offset hyperparameters. Here we take $S = 50$ values uniformly sampled between 0 and threshold c. The cosine cutoff function is defined as:

$$f(r)^{\text{cutoff}} = \begin{cases} 0.5\left[1 + \cos\left(\frac{\pi r}{c}\right)\right] & r < c \\ 0 & r \geqslant c. \end{cases} \tag{3}$$

Based on this, the distance representation is initialized as $\boldsymbol{h}^{\text{dist}} = f(r)^{\text{RBF}} \cdot f(r)^{\text{cutoff}} \in \mathbb{R}^S$. With these input representations, we follow the message-passing framework [13] to update the information between node scalar representation $\boldsymbol{h}^s$ and node vector representation $\boldsymbol{h}^v$. For the l-th layer, we have:

$$\boldsymbol{h}_l^s = \boldsymbol{h}_{l-1}^s + \text{MultiHeadAttn}(\boldsymbol{h}_{l-1}^s, \boldsymbol{h}_{l-1}^v, \boldsymbol{h}^{\text{dist}}), \quad \boldsymbol{h}_l^v = \boldsymbol{h}_{l-1}^v + \text{MultiHeadAttn}(\boldsymbol{h}_{l-1}^s, \boldsymbol{h}_{l-1}^v, \boldsymbol{h}^{\text{dist}}), \tag{4}$$

where MLP is the multiple-layer perception, and $\text{MultiHeadAttn}(\cdot, \cdot)$ is the multi-head attention module [37] which captures different aspects or types of relationships between atoms in the Euclidean space simultaneously. Below, we detail the multi-head attention architecture.

Details of Multi-Head Attention. We introduce the following multi-head attention mechanics to enable the message passing between scalar- and vector-level node representations.

For both the scalar and vector node representations, we have three matrices: query, key, and value matrix, respectively. For notation, we have $W_Q^s, W_Q^v \in \mathbb{R}^{d \times d_Q}$, $W_K^s, W_Q^v \in \mathbb{R}^{d \times d_K}$, $W_V^s, W_V^v \in \mathbb{R}^{d \times d_V}$, where d_Q, d_K, and d_V delegate the projection dimension for each matrix. Here we consider $d_Q = d_K = d$. This gives us the attention score for each atom pair within cutoff c between scalar and vector representations as:

$$\begin{aligned}
\alpha_{ij}^{s\to s} &= \text{FO}(\mathcal{F}_i \cdot \mathcal{F}_j) + \text{MLP}^{s\to s}(\boldsymbol{h}^{\text{dist}}) + \frac{\big(\boldsymbol{h}_i^s W_Q^s\big)\big(\boldsymbol{h}_j^s W_K^s\big)^T}{\sqrt{d}} \in \mathbb{R}, \\
\alpha_{ij}^{v\to s} &= \text{FO}(\mathcal{F}_i \cdot \mathcal{F}_j) + \text{MLP}^{v\to s}(\boldsymbol{h}^{\text{dist}}) + \frac{\big(\text{proj}_{\mathcal{F}_i} \boldsymbol{h}_i^v W_Q^v\big)\big(\boldsymbol{h}_j^s W_K^s\big)^T}{\sqrt{d}} \in \mathbb{R}, \\
\alpha_{ij}^{v\to v} &= \text{FO}(\mathcal{F}_i \cdot \mathcal{F}_j) + \text{MLP}^{v\to v}(\boldsymbol{h}^{\text{dist}}) + \frac{\big(\text{proj}_{\mathcal{F}_i} \boldsymbol{h}_i^v W_Q^v\big)\big(\text{proj}_{\mathcal{F}_j} \boldsymbol{h}_j^v W_K^v\big)^T}{\sqrt{d}} \in \mathbb{R}, \\
\alpha_{ij}^{s\to v} &= \text{FO}(\mathcal{F}_i \cdot \mathcal{F}_j) + \text{MLP}^{s\to v}(\boldsymbol{h}^{\text{dist}}) + \frac{\big(\boldsymbol{h}_i^s W_Q^s\big)\big(\text{proj}_{\mathcal{F}_j} \boldsymbol{h}_j^v W_K^v\big)^T}{\sqrt{d}} \in \mathbb{R},
\end{aligned} \tag{5}$$

where $\text{FO}(\mathcal{F}_i, \mathcal{F}_j)$ is the frame orientation between each pair of atom-level vector frames, defined as:

$$\text{FO}(\mathcal{F}_i, \mathcal{F}_j) = \text{MLP}(\mathcal{F}_i, \mathcal{F}_j) \in \mathbb{R}, \tag{6}$$

and the $\text{proj}_{\mathcal{F}} \boldsymbol{h}^v$ is defined as the projection from a vector representation $\boldsymbol{h}^v$ to the local frame $\mathcal{F}$ followed with an MLP layer:

$$\text{proj}_{\mathcal{F}} \boldsymbol{h}^v = \text{MLP}(\mathcal{F} \cdot \boldsymbol{h}^v) \in \mathbb{R}^{1\times d}. \tag{7}$$

Utilizing the scores defined in Equation (5) enables us to define the attention head for node scalar and vector as:

$$\begin{aligned}
\text{Attn}(\boldsymbol{h}^s) &= \text{softmax}_j(\alpha_{ij}^{s\to s})\boldsymbol{h}_j^s W_V^s + \text{softmax}_j(\alpha_{ij}^{v\to s})\,\text{proj}_{\mathcal{F}_i} \boldsymbol{h}_j^v W_V^v \\
&= \frac{\exp\big(\alpha_{ij}^{s\to s}\big)}{\sum_j \exp\big(\alpha_{ij}^{s\to s}\big)}\boldsymbol{h}_j^s W_V^s + \frac{\exp\big(\alpha_{ij}^{v\to s}\big)}{\sum_j \exp\big(\alpha_{ij}^{v\to s}\big)}\,\text{proj}_{\mathcal{F}_i} \boldsymbol{h}_j^v W_V^v, \\
\text{Attn}(\boldsymbol{h}^v) &= \text{softmax}_j(\alpha_{ij}^{v\to v})\boldsymbol{h}_j^v W_V^v + \text{softmax}_j(\alpha_{ij}^{v\to s})\boldsymbol{h}_j^s W_V^v \cdot \vec{d_{ij}} \\
&= \frac{\exp\big(\alpha_{ij}^{v\to v}\big)}{\sum_j \exp\big(\alpha_{ij}^{v\to v}\big)}\boldsymbol{h}_j^v W_V^s + \frac{\exp\big(\alpha_{ij}^{v\to s}\big)}{\sum_j \exp\big(\alpha_{ij}^{v\to s}\big)}\boldsymbol{h}_j^s W_V^s \cdot \vec{d_{ij}}.
\end{aligned} \tag{8}$$

We repeat Equation (8) for M number of heads, we have the multi-head attention as:

$$\begin{aligned}
\text{MultiHeadAttn}(\boldsymbol{h}^s) &= \big[\text{Attn}(\boldsymbol{h}^s)_0 \oplus \text{Attn}(\boldsymbol{h}^s)_1 \oplus ...\big] W_O^s, \\
\text{MultiHeadAttn}(\boldsymbol{h}^v) &= \big[\text{Attn}(\boldsymbol{h}^v)_0 \oplus \text{Attn}(\boldsymbol{h}^v)_1 \oplus ...\big] W_O^v.
\end{aligned} \tag{9}$$

Thus, we can obtain the representation for the next layer as:

$$\boldsymbol{h}^s = \boldsymbol{h}^s + \text{MultiHeadAttn}(\boldsymbol{h}^s), \qquad \boldsymbol{h}^v = \boldsymbol{h}^v + \text{MultiHeadAttn}(\boldsymbol{h}^v). \tag{10}$$

Summary. The node representations in the last layer ($\boldsymbol{h}_L^s$ and $\boldsymbol{h}_L^v$) capture the intra-information of each unit cell. In the next step, we will pass them to the inter-cell model to better capture the information of the periodic crystal structure.

2.2 Inter-cell Model: FourierFrameNet

In addition to ensuring euqivariance within the intra-cell of crystalline materials, another crucial aspect of their modeling is periodicity invariance at the inter-cell level. This requires the representation to remain consistent as the lattice repeats in 3D Euclidean space. To achieve this, we devise FourierFrameNet, a model designed to more effectively capture the periodicity inherent in crystal structures.

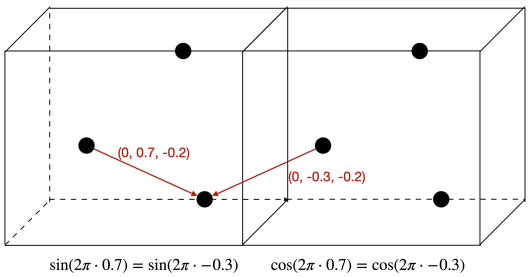

Figure 3: Illustration of periodic graph using Fourier Transform in FourierFrameNet.

Fractional coordinates are scalarization to a global vector frame. Due to the essence of scalarization, the functions merely built on fractional coordinates are invariant. The lattice is $L = [l_1, l_2, l_3]$, where $l_i \in \mathbb{E}^{1\times 3}$. Then the mapping between cartesian coordinates $\boldsymbol{x}$ and fractional coordinates $\boldsymbol{c}$ is $\boldsymbol{x} = \boldsymbol{c} \cdot L$.

The normalization enables periodic modeling of the periodic structures in crystal. For fractional coordinates, each coefficient on three axes is normalized to be between 0 and 100%.

All the interactions can be captured by using directed fully connected nodes. This is a directional (not undirectional) graph. $\tilde{\boldsymbol{x}}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)\%L \in \mathbb{E}^3$. Then, we adopt the positional encoding as follows:

$$\boldsymbol{h}_{ij}(\tilde{\boldsymbol{x}}_{ij}, f) = \sin(2\pi f \tilde{\boldsymbol{x}}_{ij}), \qquad \boldsymbol{h}_{ij}(\tilde{\boldsymbol{x}}_{ij}, f) = \cos(2\pi f \tilde{\boldsymbol{x}}_{ij}), \tag{11}$$

where $f \in \{1, 2, ..., F\}$ is the frequency. As illustrated in Figure 3, after we project the fractional coordinates into the Fourier realm, the periodicity can be captured simultaneously. This also enables capturing long-range interaction.

Multi-head Attention. We adopt a similar idea to the multi-head attention used in intra-crystal modeling (Equation (10)). Yet, we would like to add another channel describing the periodic information, *i.e.*, MLP($\boldsymbol{h}^{\text{fourier}}$) between scalars and vectors. The general pipeline is illustrated in Figure 2.

Properties of FourierFrameNet. Last but not least, we want to summarize the main attributes of FourierFrameNet. (1) For intra-cell level representation, the FrameNet is SE(3)-equivariant. For the inter-cell level representation, the FourierFrameNet is periodic invariant. (2) FourierFrameNet is agnostic to the shifting of periodic boundaries, because the atom pairwise coordinates transform equivariantly. (3) FourierFrameNet is able to handle the long-range interaction after the projection to the Fourier realm.

3 GeoSSL-PFM: Geometric Pretraining

The key idea of geometric pretraining is to learn from the crystal structure itself on a large dataset with high quality. Since only the structural information is utilized for each crystal, and no supervised signals (*e.g.*, crystal properties) are considered, this is typically called *unsupervised pretraining* or *self-supervised pretraining*.

We would like to follow the self-supervised learning paradigm for single-modal pretraining. The high-level idea is to fully explore the inherent structures in the molecules. One classical work along this line is GeoSSL [25]. The high-level idea is that the pretraining task is to maximize the mutual information (MI) between two views, and for pure geometric data like molecule conformation, we can treat the geometry provided by the dataset as the first view $\boldsymbol{g}_1$, and the perturbed geometry (*e.g.*, adding a small noise) as the second view $\boldsymbol{g}_2$. This is illustrated in Figure 1. Then we want to learn the most informative geometry between the two views, where such dependence between two variables can be measured by the mutual information:

$$\text{MI}(\boldsymbol{g}_1, \boldsymbol{g}_2). \tag{12}$$

Then following existing paradigm [25], we transform this MI maximization problem $\text{MI}(\boldsymbol{g}_1, \boldsymbol{g}_2)$ as the summation of two conditional log-likelihoods, *i.e.*,

$$\mathcal{L}_{\text{MI}} = \mathbb{E}_{\boldsymbol{g}_1}[\boldsymbol{g}_2|\boldsymbol{g}_1] + \mathbb{E}_{\boldsymbol{g}_2}[\boldsymbol{g}_1|\boldsymbol{g}_2]. \tag{13}$$

GeoSSL-InfoNCE, EBM-NCE, RR, and DDM. To estimate Equation (13), we have multiple solutions to solving this density estimation problem, including but not limited to InfoNCE [30], EBM-NCE [26], Representation Reconstruction (RR) [26], and Distance Denoising Matching (DDM) [25]. Experiments on small molecules reveal that DDM outperforms other pretraining methods in such geometric data settings. Additionally, a more recent work systematically studies the effect of different conditional generative models on crystals, including denoising diffusion and flow matching [27], and the brief conclusion is that flow matching leads to more robust performance. Thus, to this end, we propose a novel pretraining paradigm with flow matching to solve Equation (13).

GeoSSL-PFM. The novel objective is called Position Flow Matching (PFM). Following the optimal transport Gaussian path [22], the objective function in Equation (13) can be further turned as:

$$\mathcal{L}_{\text{PFM}} = \mathbb{E}_t\|\boldsymbol{g}_1 - \boldsymbol{g}_2 - v_\theta(\boldsymbol{g}_t, t)\|^2 + \mathbb{E}_t\|\boldsymbol{g}_2 - \boldsymbol{g}_1 - v_\theta(\boldsymbol{g}_t, t)\|^2. \tag{14}$$

Notice that in GeoSSL-PFM Equation (14), the velocity function outputs a velocity function in the $\mathbb{R}^3$ space, which should be SE(3)-equivariant. This contrasts with DDM, where the distance prediction only needs to be SE(3)-invariant.

4 Experiments

Pretraining dataset. Cambridge Structure Database (CSD) is the world's largest database of small-molecule organic and metal-organic crystal structure data [14], and is managed by Cambridge Crystallographic Data Center (CCDC). We utilized CSD 2023.3 for development, and adopt CSD Python API 3.1.0 to transform the crystal structures into CIF files. There are in total 1,304,168 crystal structures, and we only keep structures satisfying the following conditions: (1) Crystals with valid structures. (2) Crystals with one single component. (3) Crystals with ordered structures. (4) Crystals that are not polymers. Finally, this gives us 456,822 crystal structures, and we have 420,855 after step 2. We acknowledge that certain synthetic data techniques can be helpful to train foundation models [12, 35], and we would like to leave this for future exploration.

MatBench [10] is a test suite for benchmarking 13 machine learning model performances for predicting different material properties. The dataset size for these tasks varies from 312 to 132k. The MatBench dataset has been pre-processed to clean up the task-irrelevant and unphysical-computed data. For benchmarking, we take 8 regression tasks with crystal structure data. These tasks are [8, 11, 18] Formation energy per Perovskite cell (Per. E_{form}), Refractive index (Dielectric), Shear modulus ($log_{10}G$), Bulk modulus($log_{10}K$), exfoliation energy (E_{exfo}), frequency at last phonon PhDOS peak (Phonons), band gap (Band Gap), and formation energy (E_{form}). Detailed explanations are as follows:

- Perovskites: predicting formation energy from the crystal structure.
- Dielectric: predicting refractive index from the crystal structure.
- $log_{10}G$: predicting DFT log10 VRH-average shear modulus from crystal structure.
- $log_{10}K$: predicting DFT log10 VRH-average bulk modulus from crystal structure.
- E_{exfo}: predicting exfoliation energies from the crystal structure.
- Phonons: predicting vibration properties from the crystal structure.
- Band Gap: predicting DFT PBE band gap from the crystal structure.
- E_{form}: predicting DFT formation energy from the crystal structure.

The unit for each task is listed in Table 2.

Results. The dataset size for each task is listed above. For benchmarking, we take 60%-20%-20% as training-validation-testing for all tasks. As observed in Table 1, GeoSSL-PFM reaches the optimal results on six out of eight MatBench tasks, while being very competitive on the remaining two tasks.

Table 1: Results on the 8 tasks from MatBench. The backbone model is FourierFrameNet. The data split and task unit are in Section 4, and the metric is the mean absolute error (MAE). The optimal results are **bolded**.

Model	Per. E_{form} ↓ 18,928	Dielectric ↓ 4,764	$log_{10}G$ ↓ 10,987	$log_{10}K$ ↓ 10,987	E_{exfo} ↓ 636	Phonons ↓ 1,265	Band Gap ↓ 106,113	E_{form} ↓ 132,752
Random Init	0.035	0.287	0.082	0.060	67.635	46.693	**0.214**	0.035
GeoSSL-RR	0.036	0.291	0.083	0.061	70.367	63.267	0.227	0.036
GeoSSL-EBM-NCE	0.035	0.297	0.085	0.061	67.914	49.221	0.240	0.036
GeoSSL-InfoNCE	0.035	0.301	0.083	0.062	63.516	44.940	0.237	0.040
GeoSSL-DDM	**0.033**	**0.284**	0.079	0.061	61.651	45.920	0.228	0.036
GeoSSL-PFM	**0.033**	0.304	**0.077**	**0.059**	**56.098**	**38.181**	0.222	**0.034**

Table 2: Unit, dataset size, and naming specifications for MatBench.

Column in MatBench	Perovskites	Dielectric	log gvrh	log kvrh	jdft2d	Phonons	Band Gap	E Form
Task Name in Table 1	Per. E_{form}	Dielectric	$log_{10}G$	$log_{10}K$	E_{exfo}	Phonons	Band Gap	E_{form}
Size	18,928	4,764	10,987	10,987	636	1,265	106,113	132,752
Unit	eV	–	log_{10} GPa	log_{10} GPa	meV	cm^{-1}	eV	eV/atom

5 Conclusion

In this paper, we proposed NeuralCrystal, a geometric foundation model for material discovery. We first introduce a key FourierFrameNet to handle the SE(3)-equivariance and periodicity invariance in crystalline materials. We then propose GeoSSL-PFM, an SE(3)-equivariant flow matching framework

to denoise the small perturbation on the geometric data from CSD. Empirical results on eight property prediction tasks verify the effectiveness of NeuralCrystal.

Acknowledgement

This research partially used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility using NERSC award NERSC DDR-ERCAP0031157.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 12

[2] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340*, 2023. 1

[3] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023. 1, 2

[4] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021. 1, 10

[5] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 10

[6] Martin Dietrich Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2000. 3

[7] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016. 10

[8] Pierre-Paul De Breuck, Matthew L Evans, and Gian-Marco Rignanese. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on modnet. *Journal of Physics: Condensed Matter*, 33(40):404002, 2021. 6

[9] Weitao Du, Jiujiu Chen, Xuecang Zhang, Zhiming Ma, and Shengchao Liu. Molecule joint auto-encoding: trajectory pretraining with 2d and 3d diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 55077–55096, 2023. 2

[10] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *arXiv.org*, 6, 2020. 6

[11] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020. 6

[12] Xiaojing Fan and Chunliang Tao. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness. *arXiv preprint arXiv:2408.04585*, 2024. 6

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017. 2, 3

[14] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016. 2, 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 11

[16] 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. 1

[17] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024. 1

[18] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013. 6

[19] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 11

[21] Kin Long Kelvin Lee, Carmelo Gonzales, Matthew Spellings, Mikhail Galkin, Santiago Miret, and Nalini Kumar. Towards foundation models for materials science: The open matsci ml toolkit. In *Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, pages 51–59, 2023. 1

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[23] Shengchao Liu, Weitao Du, Yanjing Li, Zhuoxinran Li, Zhiling Zheng, Chenru Duan, Zhi-Ming Ma, Omar M. Yaghi, Anima Anandkumar, Christian Borgs, Jennifer T Chayes, Hongyu Guo, and Jian Tang. Symmetry-informed geometric representation for molecules, proteins, and crystalline materials. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 2, 3, 10

[24] Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pages 21497–21526. PMLR, 2023. 2

[25] Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *ICLR*, 2023. 5

[26] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *ICLR*, 2022. 2, 5

[27] Shengchao Liu, Divin Yan, Hongyu Guo, and Anima Anandkumar. Equivariant flow matching framework for learning molecular cluster crystallization. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2024. 5

[28] Shengchao Liu, Divin Yan, Hongyu Guo, and Anima Anandkumar. An equivariant flow matching framework for learning molecular crystallization. In *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2024. 1

[29] Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. Deep learning generative model for crystal structure prediction. *arXiv preprint arXiv:2403.10846*, 2024. 1

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 5

[31] Nathaniel H Park, Tiffany J Callahan, James L Hedrick, Tim Erdmann, and Sara Capponi. Leveraging chemistry foundation models to facilitate structure focused retrieval augmented generation in multi-agent workflows for catalyst and materials design. *arXiv preprint arXiv:2408.11793*, 2024. 1

[32] Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755*, 2024. 1

[33] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022. 2

[34] Nima Shoghi, Adeesh Kolluru, John R Kitchin, Zachary W Ulissi, C Lawrence Zitnick, and Brandon M Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv preprint arXiv:2310.16802*, 2023. 2

[35] Chunliang Tao, Xiaojing Fan, and Yahe Yang. Harnessing llms for api interactions: A framework for classification and synthetic data generation. *arXiv preprint arXiv:2409.11703*, 2024. 6

[36] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 1, 10

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Hermann Weyl. *Symmetry*, volume 47. Princeton University Press, 2015. 10

[39] 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. 2

[40] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024. 2

[41] Anthony Zee. *Group theory in a nutshell for physicists*, volume 17. Princeton University Press, 2016. 10

[42] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023. 1, 10

[43] Zhiling Zheng, Ali H Alawadhi, Saumil Chheda, S Ephraim Neumann, Nakul Rampal, Shengchao Liu, Ha L Nguyen, Yen-hsu Lin, Zichao Rong, J Ilja Siepmann, et al. Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned gpt models. *Journal of the American Chemical Society*, 145(51):28284–28295, 2023. 1

[44] Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170, 2023. 1

A Preliminaries

Crystal structures with periodicity.

SE(3)-equivariance. We would like to first give a brief introduction on the SE(3)-equivariance, and for more detailed discussions of SE(3)-equivariance, please check [4, 23, 36, 42].

Symmetry means the object remains invariant after certain transformations [38], and it is everywhere on Earth, such as in animals, plants, and molecules. Formally, the set of all symmetric transformations satisfies the axioms of a group. Therefore, the group theory and its representation theory are common tools to depict such physical symmetry. **Group** is a set G equipped with a group product $\times$ satisfying:

$$(1)\ \exists \boldsymbol{e} \in G,\ \boldsymbol{a}\times\boldsymbol{e}=\boldsymbol{e}\times\boldsymbol{a}, \forall \boldsymbol{a}\in G; \quad (2)\ \boldsymbol{a}\times\boldsymbol{a}^{-1}=\boldsymbol{a}^{-1}\times\boldsymbol{a}=\boldsymbol{e}; \quad (3)\ \boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c})=\boldsymbol{a}\times\boldsymbol{b}\times\boldsymbol{c}. \tag{15}$$

Group representation is a mapping from the group G to the group of linear transformations of a vector space X with dimension d (see [41] for more rigorous definition):

$$\rho_X(\cdot): G \to \mathbb{R}^{d\times d} \qquad \text{s.t.} \quad \rho(\boldsymbol{e})=1 \ \ \wedge \ \ \rho_X(\boldsymbol{a})\rho_X(\boldsymbol{b})=\rho_X(\boldsymbol{a}\times\boldsymbol{b}),\ \forall \boldsymbol{a},\boldsymbol{b}\in G. \tag{16}$$

During modeling, the X space can be the input 3D Euclidean space, the equivariant vector space in the intermediate layers, or the output force space. This enables the definition of equivariance as below.

Equivariance is the property for the geometric modeling function $f: X \to Y$ as:

$$f(\rho_X(\boldsymbol{a})\boldsymbol{x})=\rho_Y(\boldsymbol{a})f(\boldsymbol{x}),\ \ \forall \boldsymbol{a}\in G, \boldsymbol{x}\in X. \tag{17}$$

For molecule geometric modeling, the property should be rotation-equivariant and translation-equivariant (*i.e.*, SE(3)-equivariant). More concretely, $\rho_X(\boldsymbol{a})$ and $\rho_Y(\boldsymbol{a})$ are the SE(3) group representations on the input (*e.g.*, atom coordinates) and output space (*e.g.*, force space), respectively. SE(3)-equivariant modeling in Equation (17) is essentially saying that the designed deep learning model f is modeling the whole transformation trajectory on the molecule conformations, and the output is the transformed $\hat{y}$ accordingly. Further, we want to highlight that, in addition to the network architecture or representation function, the input features can also be represented as an equivariant feature mapping from the 3D mesh to $\mathbb{R}^{\tilde{d}}$ [7], where $\tilde{d}$ depends on input data, *e.g.*, $\tilde{d}$ = 1 (for atom type dimension) + 3 (for atom coordinate dimension) on small molecules. Such features are called steerable features in [5, 7] when only considering the subgroup SO(3)-equivariance.

Invariance is a special type of equivariance, defined as:

$$f(\rho_X(\boldsymbol{a})\boldsymbol{x})=f(\boldsymbol{x}),\ \ \forall \boldsymbol{a}\in G, \boldsymbol{x}\in X, \tag{18}$$

with $\rho_Y(\boldsymbol{a})$ as the identity $\forall \boldsymbol{a}\in G$. The group representation helps define the equivariance condition for f to follow. Then, the question boils down to how to design such an equivariant f. In the following, we will discuss geometric modelings from a novel and unified perspective using the frame.

Periodicity-invariance. Crystals exhibit long-range order and periodicity in their atomic or molecular arrangement. This periodic arrangement is represented by a *lattice*, which is an array of points showing the potential positions of atoms or molecules. The periodicity of the crystal structure is defined by the repeating pattern of the lattice and the specific arrangement of atoms or molecules within the lattice. In essence, the lattice describes the overall periodic pattern of the crystal structure, while periodicity refers to the regular, repeating arrangement of atoms or molecules within that lattice pattern. The crystal is the physical manifestation of this periodic lattice arrangement of atoms or molecules.

Suppose the lattice for a molecule is L. Then the periodic invariance is saying that the property is invariant when we shift the unit cell along three axes, *i.e.*, $p(y|x)=p(y|x+L)$.

A.1 Preliminaries: DDPM

First, we assume the data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, and we adopt a Markovnian noising process q that gradually adds noise to the data $\boldsymbol{x}_0$ through $\boldsymbol{x}_T$. Each noising process added a Gaussian noise by a given variance β_t:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \triangleq \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I). \tag{19}$$

Following this, we can obtain an analytical form of $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ in a Gaussian distribution. Let us define $\alpha_t \triangleq 1-\beta_t$ and $\bar{\alpha}_t \triangleq \prod_{l=0}^{t}\alpha_l$, then we have:

$$\begin{aligned} &q(\boldsymbol{x}_t|\boldsymbol{x}_0)=\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I). \\ \Longrightarrow\ &\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\epsilon_t\sqrt{1-\bar{\alpha}_t}. \end{aligned} \tag{20}$$

Then using the Bayes theorem, the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ can also be expressed as a Gaussian distribution:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t I), \tag{21}$$

where the mean and variance are:

$$\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t, \qquad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}. \tag{22}$$

By plugging in Equation (20), *i.e.*, $\sqrt{\alpha_t}\boldsymbol{x}_0 = \boldsymbol{x}_t - \epsilon_t\sqrt{1-\bar{\alpha}_t}$, we can have

$$\begin{aligned}
\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \\
&= \frac{(1-\alpha_t)(\boldsymbol{x}_t - \epsilon_t\sqrt{1-\bar{\alpha}_t})}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \\
&= \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\Big).
\end{aligned} \tag{23}$$

Parameterization with neural networks. For sampling from the data distribution $q(\boldsymbol{x}_0)$, we can first sample from $q(\boldsymbol{x}_T)$ and then follow the reverse steps $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. As proved in [15], $q(\boldsymbol{x}_T)$ is nearly isotropic Gaussian with certain settings for $\beta_T \to 0$ and $T \to \infty$. Thus, the question is how to approximate $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, and we can train neural networks to predict the mean and diagonal covariance matrix:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)). \tag{24}$$

Training objective. The objective is to maximize the log-likelihood of the estimated data distribution, that is to maximize $\log p_\theta(\boldsymbol{x}_0)$. With Jensen's inequality or by introducing the following KL-divergence:

$$\log p_\theta(\boldsymbol{x}_0) - D_{\text{KL}}\big(q(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\big) = \mathcal{L}_{\text{vlb}}, \tag{25}$$

where $\boldsymbol{z}$ is a latent variable [20]. Thus, we can see that maximizing $\log p_\theta(\boldsymbol{x}_0)$ is equivalent to maximizing the variational lower-bound, $\mathcal{L}_{\text{vlb}}$. This is equivalent to minimize $\mathcal{L}(\boldsymbol{x}_0) = -\mathcal{L}_{\text{vlb}}$, where

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}_0) &\triangleq \mathcal{L}_0 + \sum_{t=1}^{T-1}\mathcal{L}_t + \mathcal{L}_T, \\
\mathcal{L}_0 &= -\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1), \\
\mathcal{L}_t &= D_{\text{KL}}\big(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\big), \\
\mathcal{L}_T &= D_{\text{KL}}\big(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)\big).
\end{aligned} \tag{26}$$

Training objective with reparameterization. In practice, [15] found that instead of directly parameterizing $\mu_\theta(\boldsymbol{x}_t, t)$, modeling the noise ϵ is easier for optimization. This leads to the following simplified objective function:

$$\begin{aligned}
\mathcal{L}_{\text{simplified}} &= \mathbb{E}_{t\sim[1,T], \boldsymbol{x}_0\sim q(\boldsymbol{x}_0), \epsilon\sim\mathcal{N}(0,I)}\big[\|\epsilon_t - \epsilon_\theta(\boldsymbol{x}_t, t)\|^2\big] \\
&= \mathbb{E}_{t\sim[1,T], \boldsymbol{x}_0\sim q(\boldsymbol{x}_0), \epsilon\sim\mathcal{N}(0,I)}\big[\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|^2\big].
\end{aligned} \tag{27}$$

Inference. Accordingly, for sampling, we can use the following:

$$\begin{aligned}
\boldsymbol{x}_{t-1} &= \mu_\theta(\boldsymbol{x}_t, t) + \boldsymbol{z}\sigma_t, \quad \boldsymbol{z} \sim \mathcal{N}(0, I), \\
\mu_\theta(\boldsymbol{x}_t, t) &= \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t, t)\Big).
\end{aligned} \tag{28}$$

A.2 Preliminaries: SDE

The forward process is to perturb data with SDEs. Suppose $p(\boldsymbol{x}_0)$ is the data distribution and $p(\boldsymbol{x}_T)$ is the prior distribution. Such a diffusion/forward process can be modeled as the solution to an Itô SDE:

$$d\boldsymbol{x} = f(\boldsymbol{x}, t)dt + g(t)d\boldsymbol{w}, \tag{29}$$

where $\boldsymbol{w}$ is the standard Wiener process, $f(\boldsymbol{x}_t, t)$ is a drift coefficient of $\boldsymbol{x}(t)$, and $g(t)$ is the diffusion coefficient of $\boldsymbol{x}_t$.

The backward process starts with the prior distribution $\boldsymbol{x}_T \sim p(\boldsymbol{x}_T)$, and reverses the process to obtain samples $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$. Existing work [1] illustrates that the reverse of a diffusion process is also a diffusion process:

$$d\boldsymbol{x} = \big[f(\boldsymbol{x}, t) - g^2(t)\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\big]dt + g(t)d\bar{\boldsymbol{w}}, \tag{30}$$

where $\bar{\boldsymbol{w}}$ is the standard Wiener process when from T to 0, and dt is a negative timestep.

The denoising score matching [] proposes a family of density estimation methods, by training a score model to match with the score, which is the gradient of data distribution, *i.e.*, $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$. To parameterize this paradigm into the time-dependent score model $s_\theta(\boldsymbol{x}_t, t)$, we have the object to minimize:

$$\mathcal{L} = \mathbb{E}_t\big[\lambda(t)\mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{x}_0}[\|s_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2]\big], \tag{31}$$

where λ is a positive weighting function, and t is uniformly sampled from $[0, T]$.

There are many variants of SDE models under this framework. For example, the denoising score matching [] and denoising diffusion probabilistic model (DDPM) [].

VE SDE In denoising score matching, the forward process is:

$$d\boldsymbol{x} = \sqrt{\frac{d\sigma_t^2}{dt}}d\boldsymbol{w}. \tag{32}$$

This leads to exploding variance when $t \to \infty$, so this is named variance exploding (VE) SDE.

VP SDE The DDPM forward process converges to the following SDE:

$$d\boldsymbol{x} = -\frac{1}{2}\beta_t \boldsymbol{x}dt + \sqrt{\beta_t}d\boldsymbol{w}. \tag{33}$$

This yields a process with a fixed variance when the initial distribution has a unit variance, so this is called variance preserving (VP) SDE.