# CRYSTAL STRUCTURE PREDICTION BY JOINT EQUIVARIANT DIFFUSION ON LATTICES AND FRACTIONAL COORDINATES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Crystal Structure Prediction (CSP) is crucial in various scientific disciplines. Existing learning-based generative approaches seldom capture the full symmetries of the crystal structure distribution—the invariance of translation, rotation, and periodicity. In this paper, we propose DiffCSP, a novel diffusion method to learn the stable structure distribution from data, incorporating the above symmetries. To be specific, DiffCSP jointly generates the lattice and the fractional coordinates of all atoms by employing a periodic-E(3)-equivariant denoising model to better model the crystal geometry. Notably, DiffCSP leverages fractional coordinates other than traditional Cartesian coordinates to represent crystals, remarkably promoting the diffusion and the generation process of atom positions. Extensive experiments on crystal structure prediction verify the effectiveness of DiffCSP against existing learning-based counterparts.

## 1 INTRODUCTION

Crystal Structure Prediction (CSP), which returns the stable 3D structure of a compound based solely on its composition, has been a goal in physical sciences since the 1950s (Desiraju, 2002). As crystals are the foundation of various solid materials, estimating their structures in 3D space determines the physical and chemical properties that greatly influence the application to various academic and industrial sciences, such as the design of batteries and catalysis (Butler et al., 2018).

CSP is related to two well-known tasks: protein structure prediction (Jumper et al., 2021) and molecular conformation generation (Shi et al., 2021), which aim at predicting the 3D structure of a protein sequence or a molecular graph, respectively. That being said, CSP exhibits unique challenges, mainly incurred by the periodicity of the atom arrangement in crystals. To generate such type of structures, we require to not only model the distribution of the atom coordinates within every cell, but also infer how their bases (*a.k.a.* lattices) are placed in 3D space. Furthermore, the choice of the lattice is not unique owing to the periodicity, which makes CSP much more challenging.

Conventional methods towards CSP mostly apply the computationally-intensive Density Functional Theory (DFT) (Kohn & Sham, 1965) to compute the energy at each iteration, guided by optimization algorithms (such as random search (Pickard & Needs, 2011), Bayesian optimization (Yamashita et al., 2018), e.t.c.) to iteratively search for the stable state corresponding to the local minima of the energy surface (Oganov et al., 2019). Recently, machine learning methods have been developed to learn the stable structures directly from the training data based on deep generative models (Court et al., 2020; Yang et al., 2021). Although the generative methods accelerate previous DFT-based counterparts remarkably, they seldom consider the full symmetries of the crystal structure distribution in the 3D world, giving rise to poor generalization ability. From the perspective of physics, any E(3) transformation, including translation, rotation, and reflection, of the coordinates does not change the physical law and thus keeps the distribution invariant. Moreover, the aforementioned periodicity is another vital symmetry of crystals. For conciseness, we call the E(3) invariance plus periodicity as *periodic E(3) invariance*, which, unfortunately, is less explored in existing generative models. There are also other methods that apply machine learning models to replace DFT for energy prediction followed by structure optimization (Jacobsen et al., 2018; Podryabinkin et al., 2019; Cheng et al.,

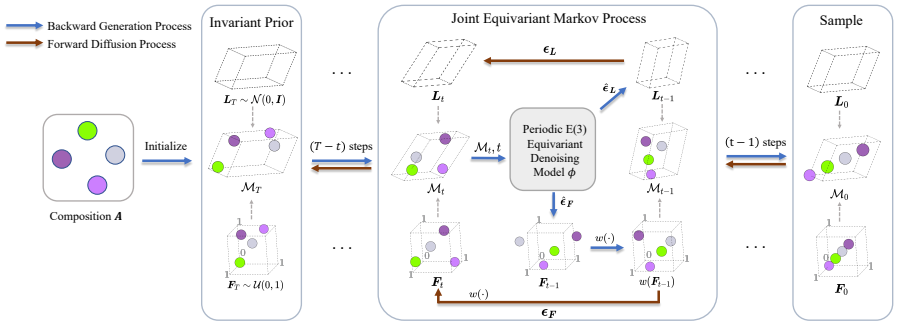


Figure 1: Overview of DiffCSP. Given the composition $\boldsymbol{A}$, we denote the crystal, its lattice and fractional coordinate matrix at time $t$ as $\mathcal{M}_t$, $\boldsymbol{L}_t$ and $\boldsymbol{F}_t$, respectively. The terms $\boldsymbol{\epsilon}_{\boldsymbol{L}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{F}}$ are Gaussian noises. $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ are predicted by the denoising model $\phi$.

2022). Although the predictors can be made E(3) invariant to reflect the symmetry, their effectiveness is still limited by the vast search space for optimization.

In this work, we introduce DiffCSP, an equivariant diffusion method to address CSP. DiffCSP is motivated by the success of diffusion models in relevant scientific domains, including molecular conformation generation (Xu et al., 2021), protein structure prediction (Trippe et al., 2022) and protein docking (Corso et al., 2022). Considering the specificity of the crystal geometry here, our DiffCSP jointly and simultaneously generates the lattice and the fractional coordinates of all atoms, by employing a proposed denoising model that is theoretically proved to be periodic E(3) invariant. A preferable characteristic of DiffCSP is that, it leverages the fractional coordinate system (defined in § 2) other than the Cartesian system used in previous methods to represent crystals, which encodes periodicity intrinsically. In particular, the fractional representation not only allows us to consider Wrapped Normal (WN) distribution (Jing et al., 2022) to better model the periodic process on fractional coordinates, but also facilitates the design of the denoising model via the Fourier transform, compared to the multi-graph encoder in crystal modeling (Xie & Grossman, 2018).

## 2 PRELIMINARIES

**Representation of crystal structures** A 3D crystal can be represented as the infinite periodic arrangement of atoms in 3D space, and the smallest repeating unit is called a *unit cell*. A unit cell can be defined by a triple $\mathcal{M} = (\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L})$, where $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_N] \in \mathbb{R}^{h \times N}$ denotes the list of the one-hot representations of atom types, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N] \in \mathbb{R}^{3 \times N}$ consists of Cartesian coordinates of the atoms, and $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3] \in \mathbb{R}^{3 \times 3}$ represents the lattice matrix containing three basic vectors describing the periodicity of the crystal. The infinite crystal structure is represented by

$$\{(\boldsymbol{a}_i', \boldsymbol{x}_i') | \boldsymbol{a}_i' = \boldsymbol{a}_i, \boldsymbol{x}_i' = \boldsymbol{x}_i + \boldsymbol{L}\boldsymbol{k}, \ \forall \boldsymbol{k} \in \mathbb{Z}^{3 \times 1}\}, \tag{1}$$

where the $j$-th element of the integral vector $\boldsymbol{k}$ denotes the integral 3D translation in units of $\boldsymbol{l}_j$.

**Fractional coordinate system** The Cartesian coordinate system $\boldsymbol{X}$ leverages three standard orthogonal bases as the coordinate axes. In crystallography, the fractional coordinate system is usually applied to reflect the periodicity of the crystal structure, which utilizes the lattices $(\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3)$ as the bases. In this way, a point represented by the fractional coordinate vector $\boldsymbol{f} = [f_1, f_2, f_3]^\top \in [0, 1)^3$ corresponds to the Cartesian vector $\boldsymbol{x} = \sum_{i=1}^{3} f_i \boldsymbol{l}_i$. This paper employs the fractional coordinate system, and denotes the crystal by $\mathcal{M} = (\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{L})$, where the fractional coordinates of all atoms compose the matrix $\boldsymbol{F} \in [0, 1)^{3 \times N}$.

**Symmetries of Crystal Structure Distribution** While various generative models can be utilized to address CSP, this task encounters particular challenges, including constraints arising from symmetries of crystal structure distribution. We formally depict the related notions below.

**Definition 1** (O(3) Invariance)**.** *For any $\boldsymbol{Q} \in \mathbb{R}^{3 \times 3}$ satisfying $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$, $p(\boldsymbol{Q}\boldsymbol{L}, \boldsymbol{F} \mid \boldsymbol{A}) = p(\boldsymbol{L}, \boldsymbol{F} \mid \boldsymbol{A})$, namely, any rotation/reflection of lattice $\boldsymbol{L}$ keeps the distribution unchanged.*

**Definition 2** (Periodic Translation Invariance)**.** *For any translation $\boldsymbol{t} \in \mathbb{R}^3$, $p(\boldsymbol{L}, w(\boldsymbol{F} + \boldsymbol{t}) \mid \boldsymbol{A}) = p(\boldsymbol{L}, \boldsymbol{F} \mid \boldsymbol{A})$, where the function $w(\boldsymbol{F}) = \boldsymbol{F} - \lfloor \boldsymbol{F} \rfloor \in [0, 1)^{3 \times N}$ returns the fractional part of each element in $\boldsymbol{F}$. It explains that any periodic translation of $\boldsymbol{F}$ will not change the distribution.*

For simplicity, we compactly term the $O(3)$ invariance and periodic translation invariance as *periodic E(3) invariance* henceforth.

**CSP task definition** The task of CSP is to predict the lattice matrix $\boldsymbol{L}$ and the fractional matrix $\boldsymbol{F}$ given the chemical composition $\boldsymbol{A}$, namely, learning the conditional distribution $p(\boldsymbol{L}, \boldsymbol{F} \mid \boldsymbol{A})$.

## 3 Overview of Proposed Method

As illustrated in Figure 1, our method DiffCSP addresses CSP by simultaneously diffusing the lattice $\boldsymbol{L}$ and the fractional coordinate matrix $\boldsymbol{F}$. Given the atom composition $\boldsymbol{A}$, $\mathcal{M}_t$ denotes the intermediate state of $\boldsymbol{L}$ and $\boldsymbol{F}$ at time step $t$ $(0 \leq t \leq T)$. DiffCSP defines two Markov processes: the forward diffusion process gradually adds noise to $\mathcal{M}_0$, and the backward generation process iteratively samples from the prior distribution $\mathcal{M}_T$ to recover the origin data $\mathcal{M}_0$.

The recovered distribution from $\mathcal{M}_T$ should meet periodic E(3) invariance. Such requirement is satisfied if the prior distribution $p(\mathcal{M}_T)$ is invariant and the Markov transition $p(\mathcal{M}_{t-1} \mid \mathcal{M}_t)$ is equivariant, according to the diffusion-based generation literature (Xu et al., 2021). Here, an equivariant transition is specified as $p(g \cdot \mathcal{M}_{t-1} \mid g \cdot \mathcal{M}_t) = p(\mathcal{M}_{t-1} \mid \mathcal{M}_t)$ where $g \cdot \mathcal{M}$ refers to any orthogonal/translational transformation $g$ acts on $\mathcal{M}$ in the way presented in Definitions 1-2. We separately summarize the derivation processes of $\boldsymbol{L}$ and $\boldsymbol{F}$ below, with more details in Appendix A.1.

**Diffusion on $\boldsymbol{L}$** Given that $\boldsymbol{L}$ is continuously variable, we exploit Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) to accomplish the generation. We utilize the forward process to progressively diffuse $\boldsymbol{L}_0$ towards the Normal prior $\mathcal{N}(0, \boldsymbol{I})$. For generation, we initialize $p(\boldsymbol{L}_T)$ from the O(3)-invariant prior distribution $\mathcal{N}(0, \boldsymbol{I})$ and apply the O(3)-equivariant backward process ensuring by the denoising model to acquire the O(3)-invariant marginal distribution $p(\boldsymbol{L}_0)$.

**Diffusion on $\boldsymbol{F}$** The domain of fractional coordinates $[0, 1)^{3\times N}$ forms a quotient space $\mathbb{R}^{3\times N}/\mathbb{Z}^{3\times N}$ induced by the crystal periodicity. It is not suitable to apply the above DDPM fashion to generate $\boldsymbol{F}$, as the normal distribution used in DDPM is unable to model the cyclical and bounded domain of $\boldsymbol{F}$. Instead, we leverage Score-Matching (SM) based framework Song & Ermon (2020); Song et al. (2020) along with Wrapped Normal (WN) distribution (De Bortoli et al., 2022) to fit the specificity here. Note that WN distribution has been explored in generative models, such as molecular conformation generation (Jing et al., 2022).

In the forward process, we first sample $\boldsymbol{\epsilon_F}$ from $\mathcal{N}(0, \boldsymbol{I})$, and then acquire $\boldsymbol{F}_t = w(\boldsymbol{F}_0 + \sigma_t \boldsymbol{\epsilon_F})$ where the truncation $w(\cdot)$ is already defined in Definition 2. This sampling implies the WN transition:

$$q(\boldsymbol{F}_t|\boldsymbol{F}_0) \propto \sum_{\boldsymbol{Z}\in\mathbb{Z}^{3\times N}} \exp\Big(-\frac{\|\boldsymbol{F}_t - \boldsymbol{F}_0 + \boldsymbol{Z}\|_F^2}{2\sigma_t^2}\Big). \tag{2}$$

Here, the noise scale $\sigma_t$ obeys the exponential scheduler. Desirably, $q(\boldsymbol{F}_t|\boldsymbol{F}_0)$ is periodic translation equivariant, and approaches a uniform distribution $\mathcal{U}(0, 1)$ if $\sigma_T$ is sufficiently large.

For the backward process, we first initialize $\boldsymbol{F}_T$ from the uniform distribution $\mathcal{U}(0, 1)$, which is periodic translation invariant. We then apply the predictor-corrector sampler (Song et al., 2020) to sample $\boldsymbol{F}_0$. The periodic translation invariance of the marginal distribution $p(\boldsymbol{F}_0)$ is further maintained by the denoising model.

**Architecture of the Denoising Model** As mentioned above, the denoising model should satisfy certain symmetries to guarantee the periodic E(3) invariance of the sampled distribution $p(\mathcal{M}_0)$. We design a message-passing neural network to model the structures and apply the inner product scalarization for O(3)-euqivariance and the Fourier transformation for periodic translation invariance. We explain the detailed architecture in Appendix A.2 along with theoretical analysis in Appendix B.

## 4 Experiments

**Dataset and metrics** We conduct experiments on three datasets with distinct levels of difficulty. **Perov-5** (Castelli et al., 2012a;b) contains 18,928 perovskite materials with similar structures. Each structure has 5 atoms in a unit cell. **MP-20** (Jain et al., 2013) selects 45,231 stable inorganic materials from Material Projects (Jain et al., 2013), which includes the majority of experimentally-generated

Table 1: Results on crystal structure prediction task.

| | # of samples | Perov-5 | | MP-20 | | MPTS-52 | |
|---|---|---|---|---|---|---|---|
| | | Match rate | RMSE | Match rate | RMSE | Match rate | RMSE |
| RS | 20 | 29.22 | 0.2924 | 8.73 | 0.2501 | 2.05 | 0.3329 |
| | 5,000 | 36.56 | 0.0886 | 11.49 | 0.2822 | 2.68 | 0.3444 |
| BO | 20 | 21.03 | 0.2830 | 8.11 | 0.2402 | 2.05 | 0.3024 |
| | 5,000 | 55.09 | 0.2037 | 12.68 | 0.2816 | 6.69 | 0.3444 |
| PSO | 20 | 20.90 | 0.0836 | 4.05 | 0.1567 | 1.06 | 0.2339 |
| | 5,000 | 21.88 | 0.0844 | 4.35 | 0.1670 | 1.09 | 0.2390 |
| P-cG-SchNet | 1 | 48.22 | 0.4179 | 15.39 | 0.3762 | 3.67 | 0.4115 |
| | 20 | 97.94 | 0.3463 | 32.64 | 0.3018 | 12.96 | 0.3942 |
| CDVAE | 1 | 45.31 | 0.1138 | 33.90 | 0.1045 | 5.34 | 0.2106 |
| | 20 | 88.51 | 0.0464 | 66.95 | 0.1026 | 20.79 | 0.2085 |
| DiffCSP | 1 | 52.02 | 0.0760 | 51.49 | 0.0631 | 12.19 | 0.1786 |
| | 20 | **98.60** | **0.0128** | **77.93** | **0.0492** | **34.02** | **0.1749** |

materials with at most 20 atoms in a unit cell. **MPTS-52** is a more challenging extension of MP-20, consisting of 40,476 structures up to 52 atoms per cell, sorted according to the earliest published year in literature. For Perov-5 and MP-20, we apply the 60-20-20 split in line with Xie et al. (2021). For MPTS-52, we split 27,380/5,000/8,096 for training/validation/testing in chronological order. For the evaluation metrics, we adopt Match rate and RMSE, with formal definitions in Appendix C.3.

**Baselines** We contrast two types of previous works. The first type follows the predict-optimize paradigm, which first trains a predictor of the target property and then utilizes certain optimization algorithms to search for optimal structures. Following Cheng et al. (2022), we apply MEGNet (Chen et al., 2019) as the predictor of the formation energy. For the optimization algorithms, we choose Random Search (**RS**), Bayesian Optimization (**BO**), and Particle Swarm Optimization (**PSO**), all iterated over 5,000 steps. The second type is based on deep generative models. We follow the modification in Xie et al. (2021) and leverage cG-SchNet (Gebauer et al., 2022) that utilizes SchNet (Schütt et al., 2018) as the backbone and additionally consider the ground-truth lattice initialization for encoding periodicity, yielding a final model named **P-cG-SchNet**. Another baseline **CDVAE** (Xie et al., 2021) is a VAE-based framework for pure crystal generation, by first predicting the lattice and the initial composition and then optimizing the atom types and coordinates via annealed Langevin dynamics (Song & Ermon, 2020). To adapt CDVAE into the CSP task, we replace the original normal prior for generation with a parametric prior conditional on the encoding of the given composition. More details are provided in Appendix C.2.

**Results** Table 1 conveys the following observations. **1.** The optimization methods encounter low Match rates, signifying the difficulty of locating the optimal structures within the vast search space. **2.** In comparison to other generative methods that construct structures atom by atom or predict the lattice and atom coordinates in two stages, our method demonstrates superior performance, highlighting the effectiveness of jointly refining the lattice and coordinates during generation. **3.** All methods struggle with performance degradation as the number of atoms per cell increases, on the datasets from Perov-5 to MPTS-52. For example, the match rates of the optimization methods are less than 10% in MPTS-52. Even so, our method consistently outperforms all other methods. More experiments on metastable structure generation and property prediction are deferred to Appendix D.1 and D.2.

## 5 Conclusion

In this work, we present DiffCSP, a diffusion-based framework for crystal structure prediction, particularly curated to take into account the vital symmetries in crystals. It is highly flexible by jointly optimizing the lattice and fractional coordinates, where the intermediate distributions are invariant under permutations, orthogonal transformations, and periodic translations. We verifies the strong applicability of DiffCSP on a wide range of crystal datasets, where it consistently matches the ground truth more closely than the baselines in terms of structural similarity and formation energy.

## REFERENCES

James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pp. 115–123. PMLR, 2013.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

Ivano E Castelli, David D Landis, Kristian S Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F Jaramillo, and Karsten W Jacobsen. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy & Environmental Science*, 5(10):9034–9043, 2012a.

Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science*, 5(2):5814–5819, 2012b.

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.

Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. Crystal structure prediction by combining graph network and optimization algorithm. *Nature communications*, 13(1):1–8, 2022.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of chemical information and modeling*, 60(10):4518–4535, 2020.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modeling. *arXiv preprint arXiv:2202.02763*, 2022.

Gautam R Desiraju. Cryptic crystallography. *Nature materials*, 1(2):77–79, 2002.

Scott Fredericks, Kevin Parrish, Dean Sayre, and Qiang Zhu. Pyxtal: A python library for crystal structure generation and symmetry analysis. *Computer Physics Communications*, 261:107810, 2021. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2020.107810. URL http://www.sciencedirect.com/science/article/pii/S0010465520304057.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Machine Learning for Molecules Workshop, NeurIPS*, 2020.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 7566–7578. Curran Associates, Inc., 2019.

Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):1–11, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

TL Jacobsen, MS Jørgensen, and B Hammer. On-the-fly machine learning of atomic potential in density functional theory structure optimization. *Physical review letters*, 120(2):026102, 2018.

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.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

Gerhard Kurz, Igor Gilitschenski, and Uwe D Hanebeck. Efficient evaluation of the probability density function of a wrapped normal distribution. In *2014 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–5. IEEE, 2014.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, 2019.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

Chris J. Pickard. Airss data for carbon at 10gpa and the c+n+h+o system at 1gpa, 2020. URL `https://archive.materialscloud.org/record/2020.0026/v1`.

Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

Evgeny V Podryabinkin, Evgeny V Tikhonov, Alexander V Shapeev, and Artem R Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6):064114, 2019.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 2021. ISSN 2590-2385. doi: https://doi.org/10.1016/j.matt.2021.11.032. URL `https://www.cell.com/matter/fulltext/S2590-2385(21)00625-1`.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

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):241722, 2018.

Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, pp. 9558–9568. PMLR, 2021.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

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.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.120.145301`.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021.

Tomoki Yamashita, Nobuya Sato, Hiori Kino, Takashi Miyake, Koji Tsuda, and Tamio Oguchi. Crystal structure prediction accelerated by bayesian optimization. *Physical Review Materials*, 2(1): 013803, 2018.

Wenhui Yang, Edirisuriya M Dilanga Siriwardane, Rongzhi Dong, Yuxin Li, and Jianjun Hu. Crystal structure prediction of materials with high symmetry using differential evolution. *Journal of Physics: Condensed Matter*, 33(45):455902, 2021.

Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10(10): 6063–6081, 2020.

# A DETAILS OF PROPOSED METHOD

In this section, we first present the joint equivaraint diffusion process on $\boldsymbol{L}$ and $\boldsymbol{F}$, and then introduce the architecture of the denoising function used in our method.

## A.1 JOINT EQUIVARIANT DIFFUSION

Algorithm 1 summarizes the forward diffusion process as well as the training of the denoising model $\phi$, while Algorithm 2 illustrates the backward sampling process. They can maintain the symmetries if $\phi$ is delicately constructed. We separately explain the derivation details of $\boldsymbol{L}$ and $\boldsymbol{F}$ below.

**Diffusion on $\boldsymbol{L}$** We first define the forward process that progressively diffuses $\boldsymbol{L}_0$ towards the Normal prior $p(\boldsymbol{L}_T) = \mathcal{N}(0, \boldsymbol{I})$ as follows:

$$q(\boldsymbol{L}_t|\boldsymbol{L}_{t-1}) = \mathcal{N}\Big(\boldsymbol{L}_t|\sqrt{1-\beta_t}\boldsymbol{L}_{t-1}, \beta_t \boldsymbol{I}\Big), \tag{3}$$

where $\beta_t \in (0, 1)$ controls the variance of the diffusion process on $\boldsymbol{L}_t$. Eq. 3 can be devised as the probability conditional on the initial distribution:

$$q(\boldsymbol{L}_t|\boldsymbol{L}_0) = \mathcal{N}\Big(\boldsymbol{L}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{L}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\Big), \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t = \prod_{s=1}^{t}(1-\beta_t)$, is valued in accordance to the cosine scheduler (Nichol & Dhariwal, 2021).

The backward generation process is given by:

$$p(\boldsymbol{L}_{t-1}|\boldsymbol{M}_t) = \mathcal{N}(\boldsymbol{L}_{t-1}|\mu(\mathcal{M}_t), \sigma^2(\mathcal{M}_t)\boldsymbol{I}), \tag{5}$$

where $\mu(\mathcal{M}_t) = \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{L}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t)\Big), \sigma^2(\mathcal{M}_t) = \beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$. The denoising term $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t) \in \mathbb{R}^{3\times 3}$ is predicted by the model $\phi(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)$.

As the prior distribution $p(\boldsymbol{L}_T) = \mathcal{N}(0, \boldsymbol{I})$ is already O(3)-invariant, we require the generation process in Eq. 5 to be O(3)-equivariant, which is formally stated below.

**Proposition 1.** *The marginal distribution $p(\boldsymbol{L}_0)$ by Algorithm 2 is O(3)-invariant if $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t)$ is O(3)-equivariant.*

To train the denoising model $\phi$, we first sample $\boldsymbol{\epsilon}_{\boldsymbol{L}} \sim \mathcal{N}(0, \boldsymbol{I})$ and reparameterize $\boldsymbol{L}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{L}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{L}}$ based on Eq. (4). The training objective is defined as the expected $\ell_2$ loss between $\boldsymbol{\epsilon}_{\boldsymbol{L}}$ and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$:

$$\mathcal{L}_{\boldsymbol{L}} = \mathbb{E}_{\boldsymbol{\epsilon}_{\boldsymbol{L}}\sim\mathcal{N}(0,\boldsymbol{I}), t\sim\mathcal{U}(1,T)}[\|\boldsymbol{\epsilon}_{\boldsymbol{L}} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t)\|_2^2]. \tag{6}$$

**Diffusion on $\boldsymbol{F}$** During the forward process, we first sample $\boldsymbol{\epsilon}_{\boldsymbol{F}}$ from $\mathcal{N}(0, \boldsymbol{I})$, and then acquire $\boldsymbol{F}_t = w(\boldsymbol{F}_0 + \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{F}})$ where the truncation function $w(\cdot)$ is already defined in Definition 2. This truncated sampling implies the WN transition:

$$q(\boldsymbol{F}_t|\boldsymbol{F}_0) \propto \sum_{\boldsymbol{Z}\in\mathbb{Z}^{3\times N}} \exp\Big(-\frac{\|\boldsymbol{F}_t - \boldsymbol{F}_0 + \boldsymbol{Z}\|_F^2}{2\sigma_t^2}\Big). \tag{7}$$

Here, the noise scale $\sigma_t$ obeys the exponential scheduler: $\sigma_0 = 0$ and $\sigma_t = \sigma_1(\frac{\sigma_T}{\sigma_1})^{\frac{t-1}{T-1}}$, if $t > 0$. Desirably, $q(\boldsymbol{F}_t|\boldsymbol{F}_0)$ is periodic translation equivariant, and approaches a uniform distribution $\mathcal{U}(0, 1)$ if $\sigma_T$ is sufficiently large.

For the backward process, we first initialize $\boldsymbol{F}_T$ from the uniform distribution $\mathcal{U}(0, 1)$, which is periodic translation invariant. We then apply the predictor-corrector sampler (Song et al., 2020) to sample $\boldsymbol{F}_0$. In Algorithm 2, Line 7 refers to the predictor while Lines 8-10 correspond to the corrector, where the term $\hat{\epsilon}_{\boldsymbol{F}} \in \mathbb{R}^{3\times N}$ is the predicted score by $\phi$. We immediately have the following proposition.

**Proposition 2.** *The marginal distribution $p(\boldsymbol{F}_0)$ by Algorithm 2 is periodic translation invariant if $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\mathcal{M}_t, t)$ is periodic translation invariant.*

**Algorithm 1** Training Procedure of DiffCSP

**Input:** lattice matrix $\boldsymbol{L}_0$, atom types $\boldsymbol{A}$, fractional coordinates $\boldsymbol{F}_0$, denoising model $\phi$, and the number of sampling steps $T$.
Sample $\boldsymbol{\epsilon_L} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$,$\boldsymbol{\epsilon_F} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $t \sim \mathcal{U}(1, T)$.
$\boldsymbol{L}_t \leftarrow \sqrt{\bar{\alpha}_t}\boldsymbol{L}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon_L}$
$\boldsymbol{F}_t \leftarrow w(\boldsymbol{F}_0 + \sigma_t \boldsymbol{\epsilon_F})$
$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}, \hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}} \leftarrow \phi(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)$
$\mathcal{L}_{\boldsymbol{L}} \leftarrow \|\boldsymbol{\epsilon_L} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}\|_2^2$
$\mathcal{L}_{\boldsymbol{F}} \leftarrow \lambda_t \|\nabla_{\boldsymbol{F}_t} \log q(\boldsymbol{F}_t|\boldsymbol{F}_0) - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}\|_2^2$
Minimize $\mathcal{L}_{\boldsymbol{L}} + \mathcal{L}_{\boldsymbol{F}}$

**Algorithm 2** Sampling Procedure of DiffCSP

**Input:** atom types $\boldsymbol{A}$, denoising model $\phi$, number of sampling steps $T$, step size of Langevin dynamics $\gamma$.
Sample $\boldsymbol{L}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$,$\boldsymbol{F}_T \sim \mathcal{U}(0, 1)$.
**for** $t \leftarrow T, \cdots, 1$ **do**
  Sample $\boldsymbol{\epsilon_L}, \boldsymbol{\epsilon_F}, \boldsymbol{\epsilon}'_{\boldsymbol{F}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}, \hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}} \leftarrow \phi(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)$.
  $\boldsymbol{L}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{L}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}) + \sqrt{\beta_t \cdot \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\boldsymbol{\epsilon_L}$.
  $\boldsymbol{F}_{t-\frac{1}{2}} \leftarrow w(\boldsymbol{F}_t + (\sigma_t^2 - \sigma_{t-1}^2)\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}} + \frac{\sigma_{t-1}\sqrt{\sigma_t^2 - \sigma_{t-1}^2}}{\sigma_t}\boldsymbol{\epsilon_F})$
  $_, \hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}} \leftarrow \phi(\boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}, t-1)$.
  $d_t \leftarrow \gamma\sigma_{t-1}/\sigma_1$
  $\boldsymbol{F}_{t-1} \leftarrow w(\boldsymbol{F}_{t-\frac{1}{2}} + d_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}} + \sqrt{2d_t}\boldsymbol{\epsilon}'_{\boldsymbol{F}})$.
**end for**
**Return** $\boldsymbol{L}_0, \boldsymbol{F}_0$.

The training objective for score matching is:

$$\mathcal{L}_{\boldsymbol{F}} = \mathbb{E}_{\boldsymbol{F}_t \sim q(\boldsymbol{F}_t|\boldsymbol{F}_0), t \sim \mathcal{U}(1,T)} \big[\lambda_t \|\nabla_{\boldsymbol{F}_t} \log q(\boldsymbol{F}_t|\boldsymbol{F}_0) - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\mathcal{M}_t, t)\|_2^2\big], \tag{8}$$

where $\lambda_t = \mathbb{E}_{\boldsymbol{F}_t}^{-1}\big[\|\nabla_{\boldsymbol{F}_t} \log q(\boldsymbol{F}_t|\boldsymbol{F}_0)\|_2^2\big]$ is approximated via Monte-Carlo sampling. More details are deferred to Appendix C.1.

### A.2 The Architecture of the Denoising Model

This subsection designs the denoising model $\phi(\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t)$ that outputs $\hat{\epsilon}_{\boldsymbol{L}}$ and $\hat{\epsilon}_{\boldsymbol{F}}$ satisfying the properties stated in Proposition 1 and 2.

Let $\boldsymbol{H}^{(s)} = [\boldsymbol{h}_1^{(s)}, \cdots, \boldsymbol{h}_N^{(s)}]$ denote the node representations of the $s$-th layer. The input feature is given by $\boldsymbol{h}_i^{(0)} = \rho(f_{\text{atom}}(\boldsymbol{a}_i), f_{\text{pos}}(t))$, where $f_{\text{atom}}$ and $f_{\text{pos}}$ are the atomic embedding and sinusoidal positional encoding (Vaswani et al., 2017; Ho et al., 2020), respectively; $\rho$ is a multi-layer perception (MLP).

Built upon EGNN (Satorras et al., 2021), the $s$-th layer message-passing is unfolded as follows:

$$\boldsymbol{m}_{ij}^{(s)} = \varphi_m(\boldsymbol{h}_i^{(s-1)}, \boldsymbol{h}_j^{(s-1)}, \boldsymbol{L}^\top \boldsymbol{L}, \psi_{\text{FT}}(\boldsymbol{f}_j - \boldsymbol{f}_i)), \tag{9}$$

$$\boldsymbol{m}_i^{(s)} = \sum_{j=1}^{N} \boldsymbol{m}_{ij}^{(s)}, \tag{10}$$

$$\boldsymbol{h}_i^{(s)} = \boldsymbol{h}_i^{(s-1)} + \varphi_h(\boldsymbol{h}_i^{(s-1)}, \boldsymbol{m}_i^{(s)}). \tag{11}$$

Here $\varphi_m$ and $\varphi_h$ are MLPs. The function $\psi_{\text{FT}} : (-1, 1)^3 \to [-1, 1]^{3\times K}$ is Fourier Transform of the relative fractional coordinate $\boldsymbol{f}_j - \boldsymbol{f}_i$. Specifically, suppose the input to be $\boldsymbol{f} = [f_1, f_2, f_3]^\top$, then the $c$-th row and $k$-th column of the output is calculated by $\psi_{\text{FT}}(\boldsymbol{f})[c, k] = \sin(2\pi m f_c)$, if $k = 2m$ (even); and $\psi_{\text{FT}}(\boldsymbol{f})[c, k] = \cos(2\pi m f_c)$, if $k = 2m + 1$ (odd). The transform $\psi_{\text{FT}}$ is able to extract various frequencies of all relative fractional distances that are helpful for crystal structure modeling,

and more importantly, $\psi_{\text{FT}}$ is periodic translation invariant, namely, $\psi_{\text{FT}}(w(\boldsymbol{f}_j + \boldsymbol{t}) - w(\boldsymbol{f}_i + \boldsymbol{t})) = \psi_{\text{FT}}(\boldsymbol{f}_j - \boldsymbol{f}_i)$ for any translation $\boldsymbol{t}$. The proof is provided in Appendix B.3.

After $S$ layers of message passing conducted on the fully connected graph, the lattice noise $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ is acquired by a linear combination of $\boldsymbol{L}$, with the weights given by the final layer:

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}} = \boldsymbol{L}\varphi_L\Big(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}_i^{(S)}\Big), \tag{12}$$

where $\varphi_L$ is an MLP with output shape as $3 \times 3$.

The fractional coordinate score $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ is output by:

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}[:, i] = \varphi_F(\boldsymbol{h}_i^{(S)}), \tag{13}$$

where $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}[:, i]$ defines the $i$-th column of $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$, and $\varphi_F$ is an MLP on the final representation.

The above formulation of the denoising model $\phi(\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t)$ ensures the following property.

**Proposition 3.** *The noise $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ by Eq. 12 is O(3)-equivariant, and the score $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ from Eq. 13 is periodic translation invariant. Hence, the generated distribution by DiffCSP in Algorithm 2 is periodic E(3) invariant.*

# B Theoretical Analysis

## B.1 Proof of Proposition 1

We first introduce the following definition to describe the equivariance and invariance from the perspective of distributions.

**Definition 3.** *We call a distribution $p(x)$ is G-invariant if for any transformation $g$ in the group $G$, $p(g \cdot x) = p(x)$, and a conditional distribution $p(x|c)$ is G-equivariant if $p(g \cdot x|g \cdot c) = p(x|c), \forall g \in G$.*

We then provide and prove the following lemma to capture the symmetry of the generation process.

**Lemma 1** (Xu et al. (2021))**.** *Consider the generation process $p(x_0) = p(x_T)\int p(x_{0:T-1}|x_t)dx_{1:T}$. If the prior distribution $p(x_T)$ is G-invariant and the Markov transitions $p(x_{t-1}|x_t), 0 < t \leq T$ are G-equivariant, the marginal distribution $p(x_0)$ is also G-invariant.*

*Proof.* For any $g \in G$, we have

$$\begin{aligned}
p(g \cdot x_0) &= p(g \cdot x_T)\int p(g \cdot x_{0:T-1}|g \cdot x_t)dx_{1:T} \\
&= p(g \cdot x_T)\int \prod_{t=1}^{T} p(g \cdot x_{t-1}|g \cdot x_t)dx_{1:T} \\
&= p(x_T)\int \prod_{t=1}^{T} p(g \cdot x_{t-1}|g \cdot x_t)dx_{1:T} \\
&= p(x_T)\int \prod_{t=1}^{T} p(x_{t-1}|x_t)dx_{1:T} \\
&= p(x_T)\int p(x_{0:T-1}|x_t)dx_{1:T} \\
&= p(x_0).
\end{aligned}$$

Hence, the marginal distribution $p(x_0)$ is G-invariant. □

The proposition 1 is rewritten and proved as follows.

**Proposition 1.** *The marginal distribution $p(\boldsymbol{L}_0)$ by Algorithm 2 is O(3)-invariant if $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t)$ is O(3)-equivariant.*

*Proof.* Consider the transition probability in Eq. (5), we have

$$p(\boldsymbol{L}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}) = \mathcal{N}(\boldsymbol{L}_{t-1}|a_t(\boldsymbol{L}_t - b_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)), \sigma_t^2\boldsymbol{I}),$$

where $a_t = \frac{1}{\sqrt{\alpha_t}}, b_t = \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}, \sigma_t^2 = \beta_t \cdot \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$ for simplicity, and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\mathcal{M}_t, t)$ is completed as $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)$. For any orthogonal transformation $\boldsymbol{Q} \in \mathbb{R}^{3\times3}, \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$, we have

$$\begin{aligned}
p(\boldsymbol{Q}\boldsymbol{L}_{t-1}|\boldsymbol{Q}\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}) &= \mathcal{N}(\boldsymbol{Q}\boldsymbol{L}_{t-1}|a_t(\boldsymbol{Q}\boldsymbol{L}_t - b_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{Q}\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)), \sigma_t^2\boldsymbol{I}) \\
&= \mathcal{N}(\boldsymbol{Q}\boldsymbol{L}_{t-1}|a_t(\boldsymbol{Q}\boldsymbol{L}_t - b_t\boldsymbol{Q}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)), \sigma_t^2\boldsymbol{I}) \\
&= \mathcal{N}(\boldsymbol{Q}\boldsymbol{L}_{t-1}|\boldsymbol{Q}\Big(a_t(\boldsymbol{L}_t - b_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t))\Big), \sigma_t^2\boldsymbol{I}) \\
&= \mathcal{N}(\boldsymbol{L}_{t-1}|a_t(\boldsymbol{L}_t - b_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)), \sigma_t^2\boldsymbol{I}) \\
&= p(\boldsymbol{L}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}).
\end{aligned}$$

As the transition is $O(3)$-equivariant and the prior distribution $\mathcal{N}(0, \boldsymbol{I})$ is $O(3)$-invariant, we prove that the the marginal distribution $p(\boldsymbol{L}_0)$ is $O(3)$-invariant based on lemma 1. □

### B.2 Proof of Proposition 2

Let $\mathcal{N}_w(\mu, \sigma^2\boldsymbol{I})$ denote the wrapped normal distribution with mean $\mu$, variance $\sigma^2$ and period 1. We first provide the following lemma.

**Lemma 8.** *If the denoising term $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t)$ is periodic translation invariant, and the transition probabilty can be formulated as $p(\boldsymbol{F}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}) = \mathcal{N}_w(\boldsymbol{F}_{t-1}|\boldsymbol{F}_t + u_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t), v_t^2\boldsymbol{I})$, where $u_t, v_t$ are functions of $t$, the transition is periodic translation equivariant.*

*Proof.* For any translation $\boldsymbol{t} \in \mathbb{R}^3$, we have

$$\begin{aligned}
&p(w(\boldsymbol{F}_{t-1} + \boldsymbol{t})|\boldsymbol{L}_t, w(\boldsymbol{F}_t + \boldsymbol{t}), \boldsymbol{A}) \\
=&\mathcal{N}_w(w(\boldsymbol{F}_{t-1} + \boldsymbol{t})|w(\boldsymbol{F}_t + \boldsymbol{t}) + u_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, w(\boldsymbol{F}_t + \boldsymbol{t}), \boldsymbol{A}, t), v_t^2\boldsymbol{I}) \\
=&\mathcal{N}_w(w(\boldsymbol{F}_{t-1} + \boldsymbol{t})|w(\boldsymbol{F}_t + \boldsymbol{t}) + u_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t), v_t^2\boldsymbol{I}) \\
=&\mathcal{N}_w(w(\boldsymbol{F}_{t-1} + \boldsymbol{t})|w\Big(\boldsymbol{F}_t + u_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t) + \boldsymbol{t}\Big), v_t^2\boldsymbol{I}) \\
=&\mathcal{N}_w(\boldsymbol{F}_{t-1}|\boldsymbol{F}_t + u_t\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t), v_t^2\boldsymbol{I}) \\
=&p(\boldsymbol{F}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}).
\end{aligned}$$

□

We rewrite proposition 2 as follows.

**Proposition 2.** *The marginal distribution $p(\boldsymbol{F}_0)$ by Algorithm 2 is periodic translation invariant if $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\mathcal{M}_t, t)$ is periodic translation invariant.*

*Proof.* The transition probability of the fractional coordinates during the Predictor-Corrector sampling can be formulated as

$$\begin{aligned}
p(\boldsymbol{F}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}) &= p_P(\boldsymbol{F}_{t-\frac{1}{2}}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A})p_C(\boldsymbol{F}_{t-1}|\boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}), \\
p_P(\boldsymbol{F}_{t-\frac{1}{2}}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}) &= \mathcal{N}_w(\boldsymbol{F}_{t-\frac{1}{2}}|\boldsymbol{F}_t + (\sigma_t^2 - \sigma_{t-1}^2)\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A}, t), \frac{\sigma_{t-1}^2(\sigma_t^2 - \sigma_{t-1}^2)}{\sigma_t^2}\boldsymbol{I}), \\
p_C(\boldsymbol{F}_{t-1}|\boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}) &= \mathcal{N}_w(\boldsymbol{F}_{t-\frac{1}{2}}|\boldsymbol{F}_t + \gamma\frac{\sigma_{t-1}}{\sigma_1}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}(\boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}, t-1), 2\gamma\frac{\sigma_{t-1}}{\sigma_1}\boldsymbol{I}),
\end{aligned}$$

where $p_P, p_C$ are the transitions of the predictor and corrector. According to lemma 8, both of the transitions are periodic translation equivariant. Therefore, the transition $p(\boldsymbol{F}_{t-1}|\boldsymbol{L}_t, \boldsymbol{F}_t, \boldsymbol{A})$ is periodic translation equivariant. As the prior distribution $\mathcal{U}(0, 1)$ is periodic translation invariant, we finally prove that the marginal distribution $p(\boldsymbol{F}_0)$ is periodic translation invariant based on lemma 1. □

### B.3 PROOF OF PROPOSITION 3

We rewrite proposition 3 as follows.

**Proposition 3.** *The noise $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ by Eq. 12 is O(3)-equivariant, and the score $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ from Eq. 13 is periodic translation invariant. Hence, the generated distribution by DiffCSP in Algorithm 2 is periodic E(3) invariant.*

*Proof.* We first prove the orthogonal invariance of the inner product term $\boldsymbol{L}^\top \boldsymbol{L}$. For any orthogonal transformation $\boldsymbol{Q} \in \mathbb{R}^{3\times 3}, \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$, we have

$$(\boldsymbol{Q}\boldsymbol{L})^\top (\boldsymbol{Q}\boldsymbol{L}) = \boldsymbol{L}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{L} = \boldsymbol{L}^\top \boldsymbol{I}\boldsymbol{L} = \boldsymbol{L}^\top \boldsymbol{L}.$$

For the Fourier Transformation, consider $k$ is even, we have

$$\begin{aligned}
&\psi_{\text{FT}}(w(\boldsymbol{f}_j + \boldsymbol{t}) - w(\boldsymbol{f}_i + \boldsymbol{t}))[c, k] \\
=&\sin\Big(2\pi m\big(w(f_{j,c} + t_c) - w(f_{i,c} + t_c)\big)\Big) \\
=&\sin\left(2\pi m(f_{j,c} - f_{i,c}) - 2\pi m\Big((f_{j,c} - f_{i,c}) - (w(f_{j,c} + t_c) - w(f_{i,c} + t_c))\Big)\right) \\
=&\sin(2\pi m(f_{j,c} - f_{i,c})) \\
=&\psi_{\text{FT}}(\boldsymbol{f}_j - \boldsymbol{f}_i)[c, k].
\end{aligned}$$

Similar results can be acquired as $k$ is odd. Therefore, we have $\psi_{\text{FT}}(w(\boldsymbol{f}_j+\boldsymbol{t})-w(\boldsymbol{f}_i+\boldsymbol{t})) = \psi_{\text{FT}}(\boldsymbol{f}_j - \boldsymbol{f}_i), \forall \boldsymbol{t} \in \mathbb{R}^3$, *i.e.*, the Fourier Transformation $\psi_{\text{FT}}$ is periodic translation invariant. According to the above, the message passing layers defined in Eq. (9)- (11) is periodic E(3) invariant. Hence, we can directly prove that the coordinate denoising term $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ is periodic translation invariant. Let $\hat{\boldsymbol{\epsilon}}_l(\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t) = \varphi_L\big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{h}_i^{(S)}\big)$. For the lattice denoising term $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}} = \boldsymbol{L}\hat{\boldsymbol{\epsilon}}_l$, we have

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{Q}\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t) &= \boldsymbol{Q}\boldsymbol{L}\hat{\boldsymbol{\epsilon}}_l(\boldsymbol{Q}\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t) \\
&= \boldsymbol{Q}\boldsymbol{L}\hat{\boldsymbol{\epsilon}}_l(\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t) \\
&= \boldsymbol{Q}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}(\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A}, t), \forall \boldsymbol{Q} \in \mathbb{R}^{3\times 3}, \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}.
\end{aligned}$$

Above all, $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ is O(3)-equivariant, and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ is periodic translation invariant. According to proposition 1 and 2, the generated distribution by DiffCSP in Algorithm 2 is periodic E(3) invariant. □

## C IMPLEMENTATION DETAILS

### C.1 APPROXIMATION OF THE WRAPPED NORMAL DISTRIBUTION

The Probability Density Function (PDF) of the wrapped normal distribution $\mathcal{N}_w(0, \sigma_t^2)$ is

$$\mathcal{N}_w(x|0, \sigma_t^2) = \frac{1}{\sqrt{2\pi}\sigma_t} \sum_{k=-\infty}^{\infty} \exp\Big(-\frac{(x-k)^2}{2\sigma_t^2}\Big),$$

where $x \in [0, 1)$. Because the above series is convergent, it is reasonable to approximate the infinite summation to a finite truncated summation (Kurz et al., 2014) as

$$f_{w,n}(x; 0, \sigma_t^2) = \frac{1}{\sqrt{2\pi}\sigma_t} \sum_{k=-n}^{n} \exp\Big(-\frac{(x-k)^2}{2\sigma_t^2}\Big).$$

And the logarithmic gradient of $f$ can be formulated as

$$\begin{aligned}
\nabla_x \log f_{w,n}(x; 0, \sigma_t^2) &= \nabla_x \log\left(\frac{1}{\sqrt{2\pi}\sigma_t} \sum_{k=-n}^{n} \exp\Big(-\frac{(x-k)^2}{2\sigma_t^2}\Big)\right) \\
&= \nabla_x \log\left(\sum_{k=-n}^{n} \exp\Big(-\frac{(x-k)^2}{2\sigma_t^2}\Big)\right) \\
&= \frac{\sum_{k=-n}^{n}(k-x)\exp\big(-\frac{(x-k)^2}{2\sigma_t^2}\big)}{\sigma_t^2 \sum_{k=-n}^{n}\exp\big(-\frac{(x-k)^2}{2\sigma_t^2}\big)}
\end{aligned}$$

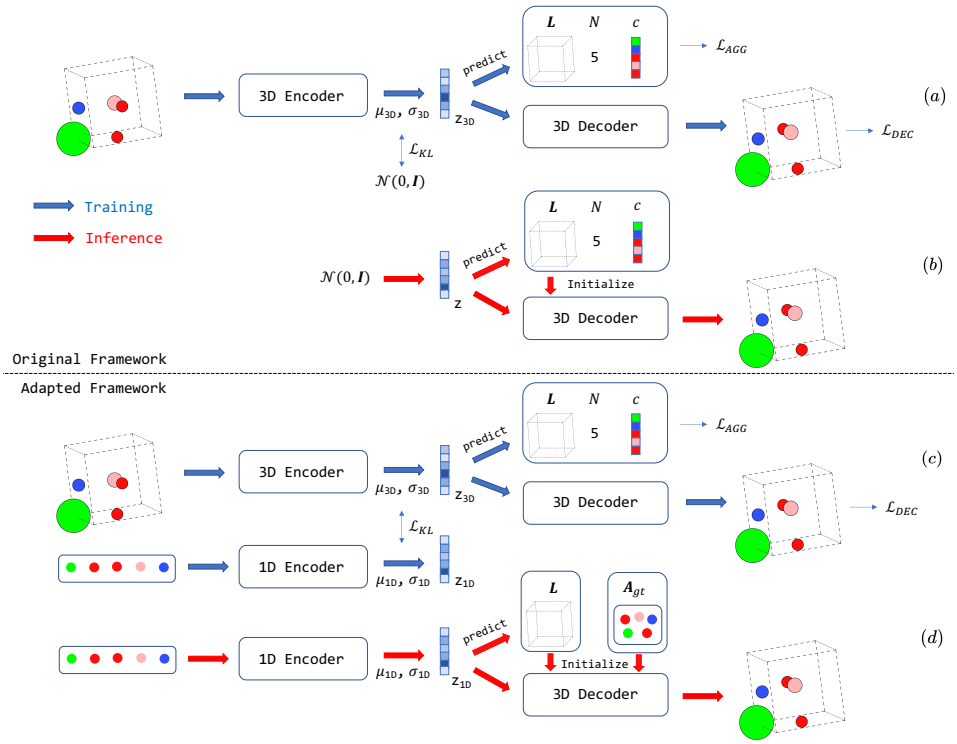


Figure 2: Overview of the original (a,b) and adapted (c,d) CDVAE. The key adaptations lie in two points. (1) We introduce an additional 1D prior encoder to fit the latent distribution of the given composition. (2) We initialize the generation procedure of the 3D decoder with the ground truth composition and keep the atom types unchanged to ensure the generated structure conforms to the given composition.

To estimate $\lambda_t = \mathbb{E}^{-1}_{x\sim\mathcal{N}_w(0,\sigma_t^2)}\left[\|\nabla_x \log \mathcal{N}_w(x|0,\sigma_t^2)\|_2^2\right]$, we first sample $m$ points from $\mathcal{N}_w(0,\sigma_t^2)$, and the expectation is approximated as

$$\begin{aligned}\tilde{\lambda}_t &= \left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_x \log f_{w,n}(x_i;0,\sigma_t^2)\|_2^2\right]^{-1} \\ &= \left[\frac{1}{m}\sum_{i=1}^{m}\left\|\frac{\sum_{k=-n}^{n}(k-x_i)\exp\left(-\frac{(x_i-k)^2}{2\sigma_t^2}\right)}{\sigma_t^2\sum_{k=-n}^{n}\exp\left(-\frac{(x_i-k)^2}{2\sigma_t^2}\right)}\right\|_2^2\right]^{-1}.\end{aligned}$$

For implementation, we select $n = 10$ and $m = 10000$.

### C.2 Adaptation of CDVAE

As illustrated in Figure 2, the original CDVAE (Xie et al., 2021) mainly consists of three parts: (1) a 3D encoder to encoder the structure into the latent variable $z_{3D}$, (2) a property predictor to predict the lattice $\boldsymbol{L}$, the number of nodes in the unit cell $N$, and the proportion of each element in the composition $c$, (3) a 3D decoder to generate the structure from $z_{3D}, \boldsymbol{L}, N, c$ via the Score Matching with Langevin Dynamics (SMLD, Song & Ermon (2020)) method. The training objective is composed of the loss functions on the three parts, *i.e.* the KL divergence between the encoded distribution and the standard normal distribution $\mathcal{L}_{KL}$, the aggregated prediction loss $\mathcal{L}_{AGG}$ and the denoising loss on the decoder $\mathcal{L}_{DEC}$. Formally, we have

$$\mathcal{L}_{ORI} = \mathcal{L}_{AGG} + \mathcal{L}_{DEC} + \beta D_{KL}\Big(\mathcal{N}(\mu_{3D}, \sigma_{3D}^2\boldsymbol{I})\|\mathcal{N}(0,\boldsymbol{I})\Big).$$

We formulate $\mathcal{L}_{KL} = \beta D_{KL}(\mathcal{N}(\mu_{3D}, \sigma_{3D}^2\boldsymbol{I})\|\mathcal{N}(0,\boldsymbol{I}))$ for better comparison with the adapted method. $\beta$ is the hyper-parameter to balance the scale of the KL divergence and other loss functions.

To adapt the CDVAE framework to the CSP task, we apply two main changes. Firstly, for the encoder side, to take the composition as the condition, we apply an additional 1D prior encoder to encode the composition set into a latent distribution $\mathcal{N}(\mu_{1D}, \sigma_{1D}^2 \boldsymbol{I})$ and minimize the KL divergence between the 3D and 1D distribution. The training objective is modified into

$$\mathcal{L}_{ADA} = \mathcal{L}_{AGG} + \mathcal{L}_{DEC} + \beta D_{KL}\Big(\mathcal{N}(\mu_{3D}, \sigma_{3D}^2 \boldsymbol{I}) \| \mathcal{N}(\mu_{1D}, \sigma_{1D}^2 \boldsymbol{I})\Big).$$

During the inference procedure, as the composition is given, the latent variable $z_{1D}$ is sampled from $\mathcal{N}(\mu_{1D}, \sigma_{1D}^2 \boldsymbol{I})$. For implementation, we apply a Transformer (Vaswani et al., 2017) without positional encoding as the 1D encoder to ensure the permutation invariance. Secondly, for the generation procedure, we apply the ground truth composition for initialization and keep the atom types unchanged during the Langevin dynamics to ensure the generated structure conforms to the given composition.

### C.3 Hyper-parameters and Training Details

We acquire the origin datasets from CDVAE (Xie et al., 2021)[1] and MPTS-52 (Jain et al., 2013)[2]. We utilize the codebases from GN-OA (Cheng et al., 2022)[3], cG-SchNet (Gebauer et al., 2022)[4] and CDVAE (Xie et al., 2021)[5] for baseline implementations.

For the optimization methods, we apply the MEGNet (Chen et al., 2019) with 3 layers, 32 hidden states as property predictor. The model is trained for 1000 epochs with an Adam optimizer with learning rate $1 \times 10^{-3}$. As for the optimization algorithms, we apply RS, PSO, and BO according to Cheng et al. (2022). For RS and BO, We employ random search and TPE-based BO as implemented in Hyperopt Bergstra et al. (2013)[6]. Specifically, we choose observation quantile $\gamma$ as 0.25 and the number of initial random points as 200 for BO. For PSO, we used scikit-opt[7] and choose the momentum parameter $\omega$ as 0.8, the cognitive as 0.5, the social parameters as 0.5 and the size of population as 20.

For P-cG-SchNet, we apply the SchNet (Schütt et al., 2018) with 9 layers, 128 hidden states as the backbone model. The model is trained for 500 epochs on each dataset with an Adam optimizer with initial learning rate $1 \times 10^{-4}$ and a Plateau scheduler with a decaying factor 0.5 and a patience of 10 epochs. We select the element proportion and the number of atoms in a unit cell as conditions for the CSP task. For CDVAE, we apply the DimeNet++ (Gasteiger et al., 2020) with 4 layers, 256 hidden states as the encoder and the GemNet-T (Gasteiger et al., 2021) with 3 layers, 128 hidden states as the decoder. We further apply a Transformer (Vaswani et al., 2017) model with 2 layers, 128 hidden states as the additional prior encoder as proposed in Appendix C.2. The model is trained for 3500, 1000, 1000 epochs for Perov-5, MP-20 and MPTS-52 respectively with an Adam optimizer with initial learning rate $1 \times 10^{-3}$ and a Plateau scheduler with a decaying factor 0.6 and a patience of 30 epochs. For our DiffCSP, we utilize the setting of 4 layer, 256 hidden states for Perov-5 and 6 layer, 512 hidden states for other datasets. The dimension of the Fourier embedding is set to $k = 256$. We apply the cosine scheduler with $s = 0.008$ to control the variance of the DDPM process on $\boldsymbol{L}_t$, and an exponential scheduler with $\sigma_1 = 0.005, \sigma_T = 0.5$ to control the noise scale of the score matching process on $\boldsymbol{F}_t$. The diffusion step is set to $T = 1000$. Our model is trained for 3500, 4000, 1000, 1000 epochs for Perov-5, Carbon-24, MP-20 and MPTS-52 with the same optimizer and learning rate scheduler as CDVAE. All models are trained on one GeForce RTX 3090 GPU.

Following the common practice (Xie et al., 2021), we evaluate by matching the predicted candidates with the ground-truth structure. Specifically, for each structure in the test set, we first generate $k$ samples of the same composition and then identify the matching if at least one of the samples matches the ground truth structure, under the metric by the StructureMatcher class in pymatgen (Ong et al., 2013) with thresholds stol=0.5, angle_tol=10, ltol=0.3. The **Match rate** is the proportion of the matched structures over the test set. **RMSE** is calculated between the ground truth and the best

[1] https://github.com/txie-93/cdvae/tree/main/data
[2] https://github.com/sparks-baird/mp-time-split
[3] http://www.comates.group/links?software=gn_oa
[4] https://github.com/atomistic-machine-learning/cG-SchNet
[5] https://github.com/txie-93/cdvae
[6] https://github.com/hyperopt/hyperopt
[7] https://github.com/guofei9987/scikit-opt

matching candidate, normalized by $\sqrt[3]{V/N}$ where $V$ is the volume of the lattice, and averaged over the matched structures. For optimization methods, we select 20 structures of the lowest energy or all 5,000 structures from all iterations during testing as candidates. For generative baselines and our DiffCSP, we let $k = 1$ and $k = 20$ for evaluation.

# D MORE EXPERIMENTS

## D.1 METASTABLE STRUCTURE GENERATION

**Dataset** We carry out experiments on **Carbon-24** (Pickard, 2020), which includes 10,153 carbon materials with 6$\sim$24 atoms in a cell. Different from the datasets adopted in section 4, where most compositions have only one stable structure for reference, Carbon-24 comprises diverse structures of a given composition. By the evaluations here, we can assess the ability to generate one-to-many metastable structures that align with the diversity of crystal structures.

**Baselines** We contrast our methods against four generative methods applicable to this dataset. **FTCP** (Ren et al., 2021) is a coordinate-based method and NOT E(3)-invariant. It represents crystals as a combination of real-space and Fourier-transformed properties fed to a CNN-VAE backbone for generation. **G-SchNet** (Gebauer et al., 2019) generates structures in an autoregressive manner and **P-G-SchNet** is a variant of G-SchNet by taking periodicity into consideration. As mentioned before, **CDVAE** (Xie et al., 2021) incorporates the score matching-based decoder into the VAE framework, and we apply its official version without any modification here. Specifically for our DiffCSP, we gather the statistics of the atom numbers from the training set, then sample the number based on the pre-computed distribution similar to the method in Hoogeboom et al. (2022), which allows DiffCSP to generate structures of variable size.

Table 2: Results on metastable structure generation task on Carbon-24. The results of baseline methods are from Xie et al. (2021).

| | Validity | Coverage | | Property | |
|---|---|---|---|---|---|
| | Valid rate(%) | COV-R | COV-P | $d_\rho$ | $d_E$ |
| FTCP | 0.08 | 0.0000 | 0.0000 | 5.206 | 19.05 |
| G-SchNet | 99.94 | 0.0000 | 0.0000 | 0.9427 | 1.32 |
| P-G-SchNet | 48.39 | 0.0000 | 0.0000 | 1.533 | 134.7 |
| CDVAE | **100.00** | 0.9980 | 0.8308 | 0.1407 | 0.285 |
| DiffCSP | **100.00** | **0.9990** | **0.9835** | **0.0590** | **0.035** |

**Evaluation Metrics** Following (Xie et al., 2021), we evaluate the generation performance from three perspectives. **Validity**: The valid rate is calculated as the percentage of the generated structures with all pairwise distances larger than 0.5Å. **Coverage**: It measures the structural similarity between the testing set $\mathcal{S}_t$ and the generated samples $\mathcal{S}_g$. Specifically, letting $d(\mathcal{M}_1, \mathcal{M}_2)$ denote the $L2$ distance of the CrystalNN fingerprints (Zimmermann & Jain, 2020) of structure, the COVerage Recall (COV-R) is determined as COV-R $= \frac{1}{|\mathcal{S}_t|}|\{\mathcal{M}_i | \mathcal{M}_i \in \mathcal{S}_t, \exists \mathcal{M}_j \in \mathcal{S}_g, d(\mathcal{M}_i, \mathcal{M}_j) < \delta\}|$ where $\delta = 0.2$ is a pre-defined threshold. The coverage precision (COV-P) is defined similarly by swapping $\mathcal{S}_g, \mathcal{S}_t$. **Property statistics**: We calculate two kinds of Wasserstein distances between the generated and testing structures, in terms of the density and the formation energy that are predicted by an independent model (Xie et al., 2021), denoted as $d_\rho$ and $d_E$, individually. The validity and coverage metrics are calculated on 10,000 generated samples, and the property metrics are evaluated on a subset with 1,000 samples passing the validity test.

**Results** Table 2 displays that our method surpasses all compared methods regarding all metrics. Specifically, DiffCSP obtains higher validity and coverage precision, indicating the high quality of the generated samples, and yields better coverage recall which reflects the promising diversity of our generated structures. Furthermore, for the property metrics, the density distance $d_\rho$ is determined by the volume of the generated lattice and the formation energy $d_E$ is highly related to the atom arrangement. DiffCSP achieves much smaller distances with respect to these two metrics, which again reveals the benefit of our joint generation mechanism.

## D.2 PROPERTY PREDICTION

The structure of a crystal plays a crucial role in determining its properties. We conduct a property prediction task on Perov-5 and MP-20 to further justify the quality of the generated sam-

ples. We apply the same 60-20-20 split in section 4. For each composition in the test set, we generate 20 samples and apply an independent predictor to predict the formation energy of each sample. The predicted energies are then averaged and compared with the ground-truth energy labels. We use Mean Absolute Error (**MAE**), symmetric Mean Absolute Percentage Error (**sMAPE**), and Pearson correlation coefficient (**PCC**) to evaluate different aspects. We compare our method with **CDVAE**, the strongest baseline in the above experiments, and **Pyxtal** (Fredericks et al., 2021), a python toolkit to generate random structures of the given composition. We also provide the results of the predictor on ground-truth structures as the vanilla reference.

From Table 3, we observe DiffCSP substantially outperforms the two baselines, demonstrating its capability of producing more precise and meaningful structures. Although there is a clear gap between the predicted structures and the ground-truth ones, DiffCSP is still able to attain comparable performance to the GT model, particularly on MP-20.

Table 3: Results on property prediction tasks. GT means the prediction results on ground truth structures, which measures the performance of the predictor and serves as the lower bound of MAE and sMAPE and the upper bound of PCC.

| | Perov-5 | | | MP-20 | | |
|---|---|---|---|---|---|---|
| | MAE | sMAPE | PCC | MAE | sMAPE | PCC |
| Pyxtal | 0.4839 | 0.3629 | 0.3363 | 0.5965 | 0.5917 | 0.8185 |
| CDVAE | 0.4760 | 0.3645 | 0.3939 | 0.0631 | 0.1431 | 0.9971 |
| DiffCSP | **0.4043** | **0.3195** | **0.5781** | **0.0318** | **0.0942** | **0.9988** |
| GT | 0.0400 | 0.0446 | 0.9950 | 0.0240 | 0.0825 | 0.9992 |

### D.3 Ablation Studies

We ablate each component of DiffCSP in Table 4. We probe the following aspects. **1.** To verify the necessity of jointly updating the lattice $\boldsymbol{L}$ and fractional coordinates $\boldsymbol{F}$ in the generation procedure, we construct two variants that separate the joint update into two stages, denoted as $\boldsymbol{L} \rightarrow \boldsymbol{F}$ and $\boldsymbol{F} \rightarrow \boldsymbol{L}$. Particularly, $\boldsymbol{L} \rightarrow \boldsymbol{F}$ applies two networks to learn the reverse processes $p_{\theta_1}(\boldsymbol{L}_{0:T-1}|\boldsymbol{A}, \boldsymbol{F}_T, \boldsymbol{L}_T)$ and $p_{\theta_2}(\boldsymbol{F}_{0:T-1}|\boldsymbol{A}, \boldsymbol{F}_T, \boldsymbol{L}_0)$. During inference, we first sample $\boldsymbol{L}_T, \boldsymbol{F}_T$ from their prior distributions, acquiring $\boldsymbol{L}_0$ via $p_{\theta_1}$, and then $\boldsymbol{F}_0$ by $p_{\theta_2}$ based on $\boldsymbol{L}_0$. $\boldsymbol{F} \rightarrow \boldsymbol{L}$ is similarly executed but with the generation order of $\boldsymbol{L}_0$ and $\boldsymbol{F}_0$ exchanged. Results indicate that $\boldsymbol{L} \rightarrow \boldsymbol{F}$ performs better than the $\boldsymbol{F} \rightarrow \boldsymbol{L}$, but both are inferior to the joint update in DiffCSP, which endorses our design. **2.** Instead of applying the score matching scheme with WN, we diffuse $\boldsymbol{F}$ via the standard normal distribution $q(\boldsymbol{F}_t|\mathcal{M}_0) = \mathcal{N}\big(\boldsymbol{F}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{F}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\big)$ during the generative process similarly defined as Eq. (5). A lower match rate and higher RMSE are observed for this variant, probably due to the lack of periodic translation invariance in the marginal distribution. **3.** Our model achieves orthogonal equivariance via the inner product $\boldsymbol{L}^\top \boldsymbol{L}$ in Eq. 9. When we replace it with $\boldsymbol{L}$ in Eq. 9 and change the final output as $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}} = \varphi_L\big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{h}_i^{(S)}\big)$ in Eq. 12 to break the equivariance, the model suffers from extreme performance detriment. Only 1.66% structures are successfully matched, which obviously implies the importance of incorporating orthogonal equivariance. **4.** We adopt Fourier embeddings to capture periodicity. To investigate its effect, we remove the Fourier embeddings from the message in Eq. 9, and the match rate drops from 51.49% to 29.15%. **5.** We further change the complete graph into the multi-graph approach adopted in Xie & Grossman (2018). The multi-graph approach decreases the match rate, since the multi-graphs constructed under different intermediate structures may differ vibrantly during generation, leading to substantially higher training difficulty and lower sampling stability.

Table 4: Ablation studies on MP-20. MG: **M**ulti-**G**raph edge construction (Xie & Grossman, 2018), FT: **F**ourier-**T**ransformation proposed in § A.2.

| | Match rate (%) | RMSE |
|---|---|---|
| $\boldsymbol{L} \rightarrow \boldsymbol{F}$ | 50.03 | 0.0921 |
| $\boldsymbol{F} \rightarrow \boldsymbol{L}$ | 36.73 | 0.0838 |
| *w/o* WN | 34.09 | 0.2350 |
| *w/o* inner product | 1.66 | 0.4002 |
| *w/o* FT | 29.15 | 0.0926 |
| MG *w/* FT | 25.85 | 0.1079 |
| MG *w/o* FT | 28.05 | 0.1314 |
| DiffCSP | **51.49** | **0.0631** |

## E IMPACT OF SAMPLING NUMBERS

Figure 3 illustrates the impact of sampling numbers on the match rate. The match rate of all methods increases when sampling more candidates, and DiffCSP outperforms the baselines methods under the arbitrary number of samples.

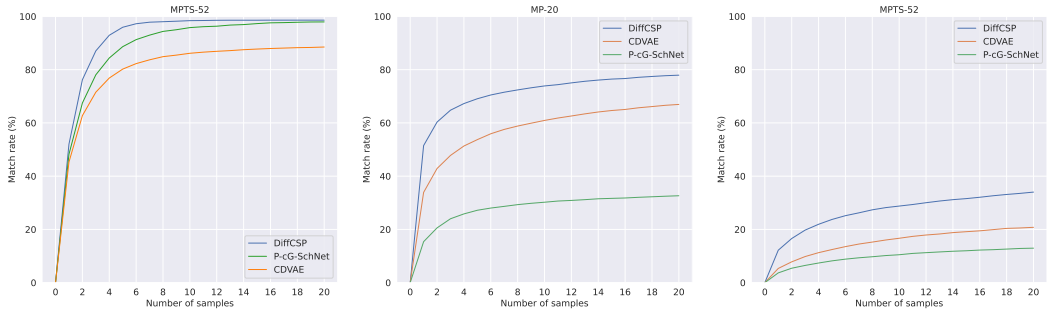


Figure 3: Comparison on different number of samples.

## F LEARNING CURVES OF DIFFERENT VARIANTS

We plot the curves of training and validation loss of different variants proposed in § D.3 in Figure 4 with the following observations. **1.** The multi-graph methods struggle with higher training and validation loss, as the edges constructed under different disturbed lattices vary significantly, complicating the training procedure. **2.** The Fourier transformation, expanding the relative coordinates and maintaining the periodic translation invariance, helps the model converge faster at the beginning of the training procedure. **3.** The variant utilizing the fully connected graph without the Fourier transformation (named "DiffCSP w/o FT" in Figure 4) encounters obvious overfitting as the periodic translation invariance is violated, highlighting the necessity to leverage the desired invariance into the model.

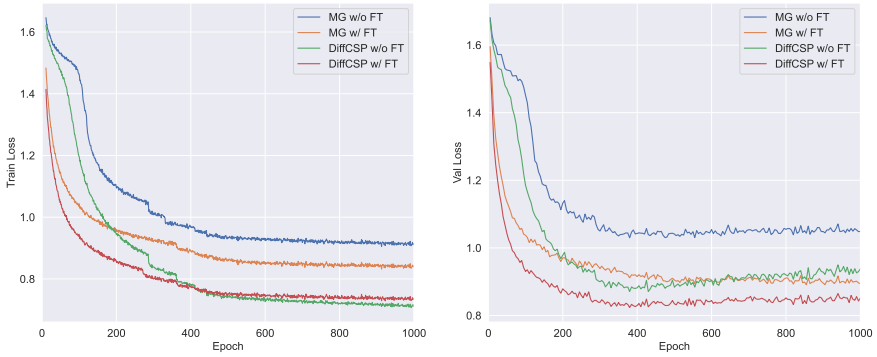


Figure 4: Learning curves of different variants proposed in § D.3. MG and FT denote multi-graph edge construction and Fourier transformation, respectively.

## G VISUALIZATIONS

In this section, we first present visualizations of the predicted structures from DiffCSP and other generative methods in Figure 5. Our DiffCSP provides more accurate predictions compared with

the baseline methods. Figure 6 illustrates 48 generated structures on Carbon-24. The visualization shows the capability of DiffCSP to generate diverse metastable structures. We further visualize the generation process in Figure 7. We find that the generated structure $\mathcal{M}_0$ is periodically translated from the ground truth structure, indicating that the marginal distribution $p(\mathcal{M}_0)$ follows the desired periodic translation invariance.

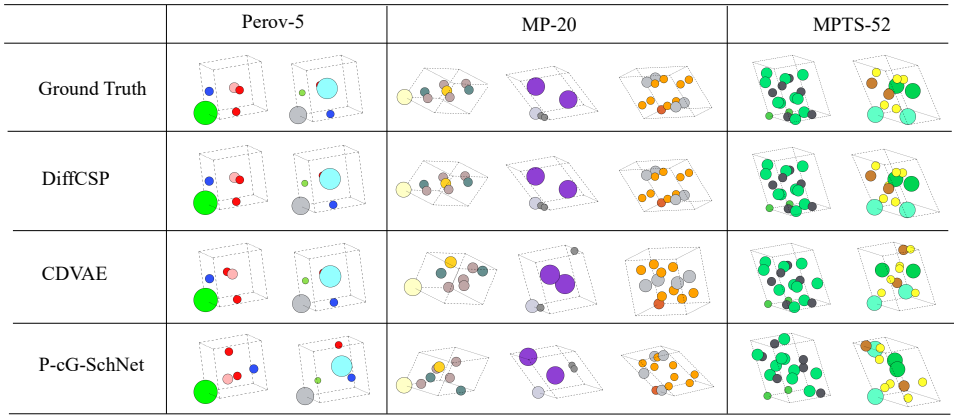


Figure 5: Additional visualizations of the predicted structures from different methods. We translate the same atom to the origin for better visualization and comparison.

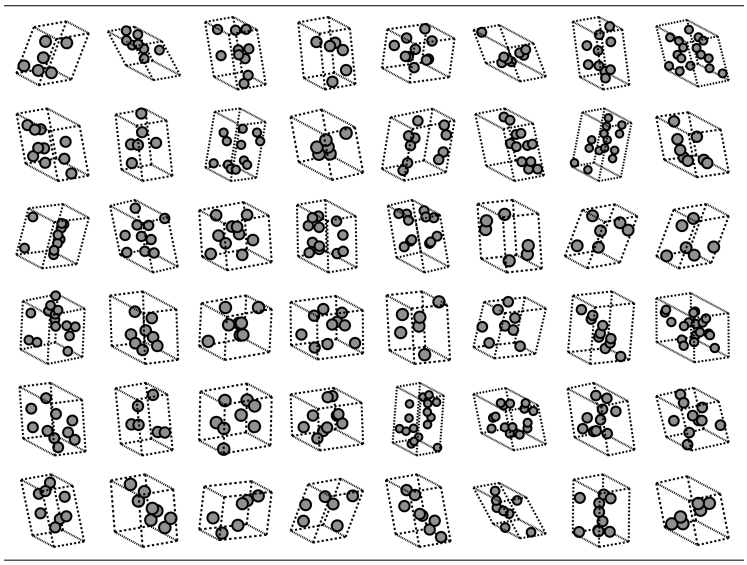

Figure 6: Visualization of the generated structures on Carbon-24.

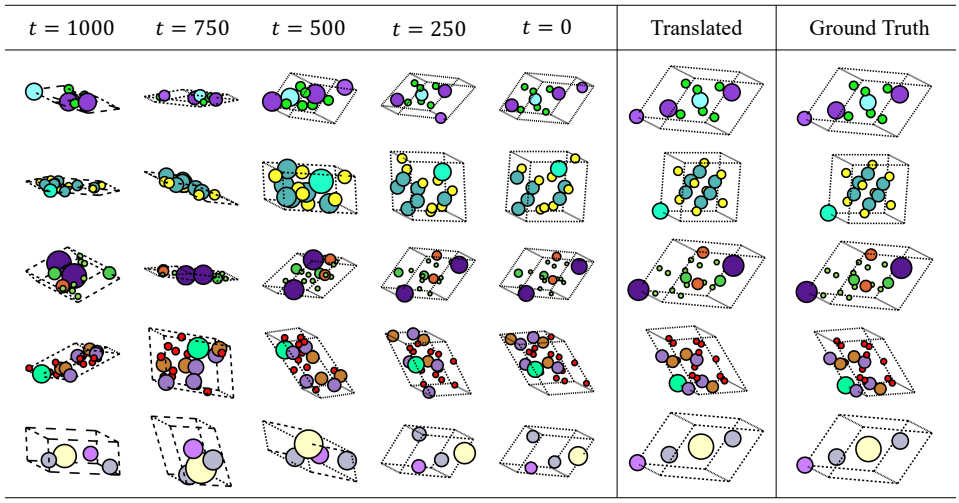


Figure 7: Visualization of the generation process on MP-20. The column "Translated" means translating the same atom in the generated structure $\mathcal{M}_0$ to the origin as the ground truth for better comparison.