DIS-CSP: DISORDERED CRYSTAL STRUCTURE PREDICTIONS

Martin H. Petersen[1,*] **Ruiming Zhu**[2] **Haiwen Dai**[2,3] **Savyasanchi Aggarwal**[3,4]
Nong Wei[2] **Andy P. Chen**[2] **Arghya Bhowmik**[1] **Juan Maria García-Lastra**[1]
Kedar Hippalgaonkar[3,4,*]
[1] Technical University of Denmark
[2] Institute of Materials Research and Engineering, ASTAR, Singapore
[3] School of Materials Science and Engineering, Nanyang Technological University, Singapore
[4] Department of Chemistry, University College London, London WC1H 0AJ, U.K.

ABSTRACT

Most synthesized crystalline inorganic materials are compositionally disordered, meaning that multiple atoms occupy the same lattice site with partial occupancy. Moreover, the computed physical properties of disordered inorganic crystals are configuration dependent, because of this partial occupancy, making it extremely challenging to solve purely by computational methods: this makes property-oriented search impractical. Crystal structure prediction (CSP), for such crystals is crucial for the eventual development of highly efficient and stable functional materials. However, existing generative models cannot handle the complexities of disordered inorganic crystals. To address this gap, we introduce an equivariant representation, based on theoretical crystallography, along with a generative model capable of generating valid structures that allow for compositional disorder and vacancies, which we call Dis-CSP. We train Dis-CSP on experimental inorganic structures from the Inorganic Crystal Structure Database (ICSD), which is the world's largest database of identified inorganic crystal structures. We show that Dis-CSP can effectively generate disordered inorganic crystal materials while preserving the inherent symmetry of the crystals throughout the generation process.

1 INTRODUCTION

The discovery of new materials plays a vital role in numerous fields of science and is crucial to developing the next generation of materials. As crystals are the foundation of various materials, crystal structure prediction (CSP) greatly influences future discovery of new materials. Traditionally, the idea of CSP is to return a 3D structure of a compound based on its composition Desiraju (2002). The ability to accurately and efficiently generate these structures paves the way for new materials discovery and design, thereby having considerable impact in many scientific fields Oganov et al. (2019). Recent advances in CSP within the inorganic crystal domain has encouraged the use of generative models for exploring this vast material space. Various strategies have been utilized such as Variational auto-encoders (VAE) Xie et al. (2022), diffusion models Jiao et al. (2023; 2024); Zeni et al. (2025); Cornet et al. (2025), transformer models Antunes et al. (2024); Kazeev et al., flow-based crystal generative models Miller et al. (2024); Sriram et al. (2024), and generative adversarial networks (GAN) Kim et al. (2020). All of these have shown promising results in the inverse design of inorganic crystalline materials.

Despite the relative success of these models, they all rely on either graph-based, or invertible feature representations of atoms occupying symmetry-defined sites in the unit cell (the repeating periodic unit) of a crystalline solid. Such representations are unfeasible when describing inorganic crystal structures with site disorder, where (1) either certain sites are occupied randomly by different types of atoms, or (2) some atomic sites are unoccupied (vacancies) causing imperfections in the crystal. Such disordered inorganic crystals are a crucial consideration, as they are very common in functional materials, with ion-mixing or doping strategies being widely applied to obtain different

properties, high-performance and stability in materials. Examples include solid solutions like the famous Mas/Oregon blue pigment $YIn_{1-x}Mn_xO_3$ Smith et al. (2009), many perovskite-structured materials with cation substitutionsNing et al. (2023); Chu et al. (2023), fast-ion conductors like α-AgI where the mobile ions occupy multiple sites within the lattice Funke et al. (2015), doped thermoelectric materials like $(PbTe)_{1-x}(PbSe)_x$ Wang et al. (2013) as well as electrodes for batteries Zhong et al. (2024); Wang et al. (2024).

To describe such ubiquitous disordered inorganic crystals computationally, it is often necessary to use multiple repetitions of the unit cell, called "supercells". This approach enables the partial occupation of atoms to be distributed among the corresponding symmetrical sites. However, due to inherent randomness of assigning the partial occupancy to the symmetrical sites, this method fails to accurately capture the true nature of the disorder in the crystal structure. Traditionally, disordered inorganic crystals are modeled with linear clustering methods, along with Monte-Carlo methods Chang et al. (2019); Su et al. (2024). This approach aims to sample the entire configurational space of the supercell, generating a representative set of supercells with various assignments of the partial occupancy to the symmetrical sites. These sampled configurations approximately describe the nature of the disordered crystal structure.

Two primary challenges associated with using this method, especially when aligned with generative models, are: (1) the need to generate thousands of supercells to describe the disordered crystal structure, and (2) the assignment of partial occupancy to specific sites inherently breaking the symmetry of the crystal structure, complicating the representation. For smaller cases, it is possible to train a generative model on a sample set of supercells, as demonstrated in Yong et al. (2024) and Zeni et al. (2025). However, generalizing such models to encompass the diverse range of inorganic disordered crystals is impractical.

A more appropriate approach is to cluster all symmetry-equivalent atomic sites in the crystal structure together, and treat each group of sites independently. This allows for the consideration of partial occupancy while preserving the site-symmetry of each group of equivalent points. These groups are referred to as Wyckoff sites. In this work, we present Dis-CSP, a framework for representing disordered inorganic crystals based on Wyckoff sites. Our framework enabling the training of a VAE model to achieve CSP on disordered crystals. This representation explicitly incorporates partial atomic occupancies by encoding both space group symmetries and Wyckoff site symmetries. In our framework, equivariance is defined empirically: any spatial operation—such as rotation or translation—that alters the Wyckoff-based encoding is effectively captured by the symmetry and reflected in model's reconstruction behavior. Additionally, the model exhibits invariance in predictive tasks, as global spatial operations do not change the physical properties or symmetry information of the crystal structure. In this paper we start by introducing the representation of disordered crystal structures, then the VAE used for generating the structures and in the reconstruction and evaluation error.

2 CRYSTAL REPRESENTATION

2.1 CRYSTAL DESCRIPTION

Crystals are highly structured solid materials defined by a repeated arrangement of atoms in space AI4Science et al. (2023). The atomic pattern that periodically repeats itself is called a motif. The parallelepiped containing the motif, which defines each periodicity in the 3D space of the motif, is called a unit cell. A $(u \times v \times w)|u, v, w \in \mathbb{N}$ repetition of the unit cell is called a supercell. Theoretical crystallography has developed methods to systemically describe the endless combinations of crystals using lattice parameters, space groups and Wyckoff sites.

A lattice Λ is an infinite set of points defined by the sum of a set of linearly independent primitive lattice vectors, $\boldsymbol{a}_i \in \mathbb{R}^n$: $\Lambda = \{R^n_{[m_i]} = \sum_i^n m_i \boldsymbol{a}_i\}$, where $m_i \in \mathbb{Z}$. In 3D, the lattice can be described by the repeated motif and 6 parameters: a, b, c determines the length of each dimension and α, β, γ determines the angle between each dimension Simon (2013).

Space groups describe the symmetry operations that the crystal can undergo while preserving the motif within the crystal lattice. In 3D, all space groups are numbered into 230 types, with the first group considered the unsymmetrical group.

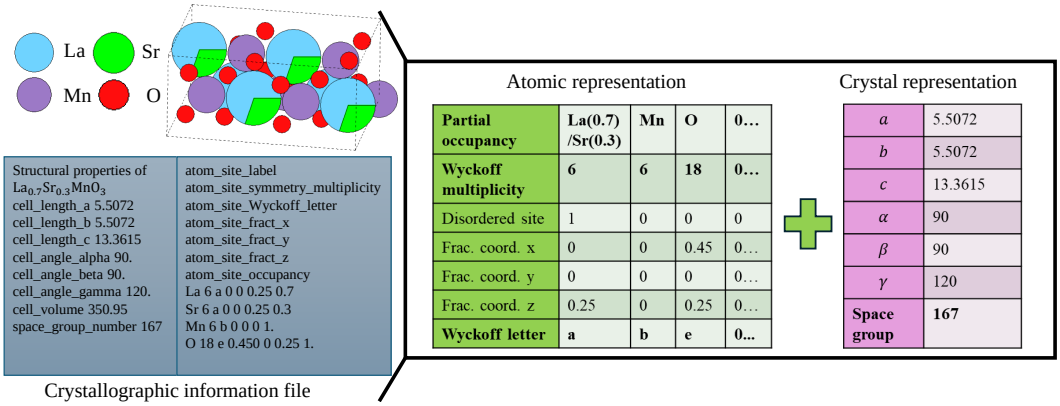


Figure 1: Structural representation of crystals. A Crystallographic information file of a crystal is transformed into a matrix $\boldsymbol{A}$ describing the atomic representation and a vector $\boldsymbol{c}$ describing the crystal representation. The bold numbers and letters represents One-hot encoding in the representation. The highlighted colors in the representations denote the labels for each rows.

For a space group $\mathbb{G}$ acting on three-dimensional space, $\mathbb{E}^3$ the (infinite) set:

$$\mathbb{O} = \mathbb{G}(X) := \{g(X) | g \in \mathbb{G}\} \tag{1}$$

is called the orbit of X under $\mathbb{G}$. The orbit of point X is the smallest subset of $\mathbb{E}^3$ that contains X and is closed under the action of $\mathbb{G}$. Every point in $\mathbb{E}^3$ belongs to exactly one orbit within the space group. The orbit partitions the direct space into disjoint subsets, meaning that an orbit is completely defined by its point in the unit cell, as translating the unit cell by the space group symmetry covers $\mathbb{E}^3$. To account for the case where two symmetry operations map X into the same point, we define subsets within the space group, termed site-symmetry groups $\mathbb{S} \in \mathbb{G}$, which define symmetry operations for X within the space group. The site-symmetry group of a point X is a finite subset of the space group, which is isomorphic to a subset of the point group $\mathcal{P}$ of the space group. The relation between site-symmetry groups of points in the same orbit is the definition of the Wyckoff site Souvignier (2015). Note that it is the definition of Wyckoff sites that any points related by symmetry operations of the space group belong to the same site. The Wyckoff sites themselves are defined using three parameters:

- Wyckoff letter, which defines the site-symmetry group for a Wyckoff site. It is labeled in alphabetical order, starting with 'a' for a position with a site-symmetry group of highest site-symmetry.
- Wyckoff multiplicity, which defines the number of points in an orbit for a Wyckoff site.
- Fractional coordinates, which define the real position in the crystal lattice for which symmetry operations can be acted upon.

Each Wyckoff site is occupied with a number of atoms equal to the Wyckoff multiplicity. If a Wyckoff site is occupied with one type of atom, it is considered ordered, and if the Wyckoff site is occupied with several types of atoms, it is disordered. In this work, an inorganic material is termed disordered if one or more Wyckoff sites are disordered and we do not consider amorphous materials in this work.

2.2 Representation of Disordered Inorganic Crystals

From the crystallography description provided in a Crystallographic Information File (CIF), a disordered inorganic crystal is represented by a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{c}$, as illustrated in Figure 1. The atomic configuration is represented by $\boldsymbol{A}$, while the crystal itself is represented by $\boldsymbol{c}$.

Each column in $\mathbf{A}$ represents the different Wyckoff sites in the crystal, while the rows describe different properties within the Wyckoff sites. The first set of rows describes the partial occupancy; the second set describes the Wyckoff multiplicity; the third set describes (1) an indicator for a disordered

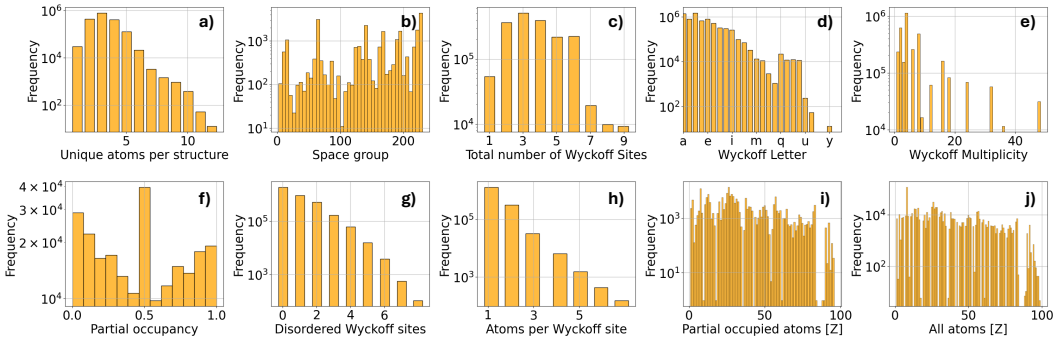


Figure 2: Data distribution of the 138,692 structures from the Inorganic Crystal Structure Database (ICSD), with 72,855 structures having partial occupied Wyckoff sites and 65,837 structures having no partial occupied Wyckoff sites. a) Number of unique atoms per structure. b) Space group per structure. c) Number of Wyckoff sites per structure. d) All Wyckoff letters. e) All Wyckoff multiplicities. f) Partial occupancy of disordered Wyckoff sites. g) Number of disordered Wyckoff sites per structure. h) Number of atoms per Wyckoff site. i) The atoms partial occupying the disordered Wyckoff sites. j) All atoms presented in the Wyckoff sites.

sites and (2) the fractional coordinates; the last set of rows describes the Wyckoff letter. The partial occupancy, Wyckoff multiplicity and Wyckoff letter are all One-hot encoded with the row number equal to their respect value, while the fractional coordinates are divided into the x, y and z components for each row. The disordered site indicator is a binary term; 1 if a Wyckoff site is disordered and 0 if not. Zero padding is added as additional columns to allow matching of crystals with fewer Wyckoff sites to those with the highest number of Wyckoff sites, such that $\boldsymbol{A}$ is the same size for all crystals. A corresponding zero padding indicator is added to the representation for the partial occupancy, the Wyckoff multiplicity and the Wyckoff letter, such that it can identify whenever a Wyckoff site exists. The first 6 entries in $\boldsymbol{c}$ consist of the 6 lattice parameters a, b, c and α, β, γ, while the last 230 entries consist of the One-hot encoded space group.

The use of Wyckoff sites and space groups to represent materials for CSP has previously been done for VAE Zhu et al. (2024) and transformer models Antunes et al. (2024); Kazeev et al.. However, to the best of our knowledge, no generalized representation has been developed to account for disordered Wyckoff sites with partial occupancy. Recently, Zeni et al. (2025) demonstrated an approach for disordered structures, but it is restricted to a specific type of disorder where two atoms swap positions during the generation of the crystal structure. Their representation does not incorporate partial occupancy of atoms, rendering an incomplete investigation of important ion-mixing and doping practices. In contrast, our representation, combined with the VAE model, provides a novel and previously unexplored framework for CSP, enabling a more comprehensive analysis of disordered crystals.

3 Disordered VAE

3.1 Data

The Inorganic Crystal Structure Database (ICSD) Hellenbrandt (2004) consists of around 229,487 precise experimental inorganic crystal structure entries, all of which are human expert inputs and verified, with 106,970 (45.6%) of the entries as disordered inorganic crystals, 122,517 (52.2%) of the entries as ordered inorganic crystals and 4966 (2.2%) of the entries having structural errors, like having the atoms to close it each other. To date, the disordered crystals of ICSD have not been utilized as training data in any generative model, instead only being used to validate generated structures as the ground truth Zhu et al. (2023); Zeni et al. (2025). Given this extensive collection of synthesized disordered inorganic crystals, the ICSD represents an ideal dataset for training generative models aimed at disordered Wyckoff sites, even in the absence of computed physical or chemical properties. To ensure smooth CSP, we reduce the complexity, common considerations for experimental crystal formation. We exclude all structures belonging to the first symmetry group

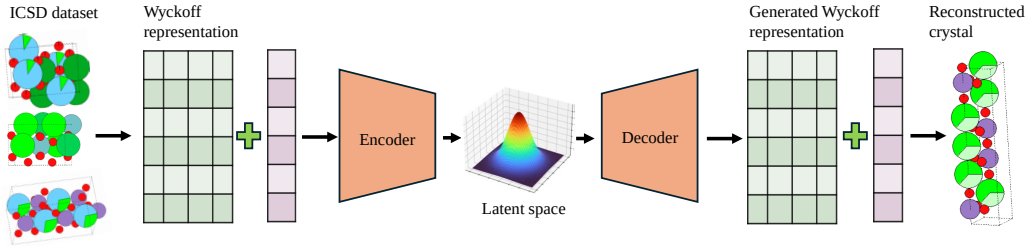


Figure 3: Schematic of the Variational auto-encoder (VAE) used in this work. All the materials gathered from ICSD are represented using the representation from Figure 1. This representation is encoded in batches and a Gaussian-based latent space is created, from which the decoding is based upon. The decoding returns a structural representation, from which the disordered structure is reconstructed.

(P1) (536 structures), since they are only assigned a unit cell without further symmetry operation check, making space group assignment incomplete. When this is the case for experimental verification of the crystal structure, we doubt that the structure itself is relaible. We also exclude structures containing rare atoms with a periodic number higher than 100 (7,515), as well as those containing more than nine Wyckoff sites (53,586 structures), more than 50 in Wyckoff multiplicity (2,756 structures), more than six disordered Wyckoff sites (25,671 structures), instances where a Wyckoff site is occupied by more than six distinct atom types (82 structures) and structures with atomic charge state as the partial occupancy (649 structures). All of these are excluded to make a dataset comparable to simpler experimental synthesis. The data distribution (138,692 structures), after filtering, is explored in Figure 2, where we see a bias towards ordered Wyckoff sites and fewer total Wyckoff sites, as expected due to the higher number of simpler crystals in the data set. In this subset, 72,855 of the crystal structures contain one or more Wyckoff sites with partial occupancy, while 65,837 of the crystal structures have no Wyckoff sites with partial disorder.

The bias toward crystals with fewer total Wyckoff sites justifies our decision to reduce the data set based on the total number of Wyckoff sites. Additionally, the use of extra zero padding would increase the complexity of the representation, which is not worthwhile since it would only accommodate a minor subset of crystals - this argument also holds for most of the other choices.

ICSD is the highest quality inorganic crystal structure database (experimental) up to date, with all information being human expert input and verified. The dataset enabled us to generate structures closer to experimental observations since we train on unit cell parameters and site occupancies experimentally measured instead of density functional theory (A) generated. All existing generative models Zeni et al. (2025); Xie et al. (2022); Jiao et al. (2023; 2024) to date assume site occupancy = 1 (which is false in many Materials Project **?** entries), using DFT unit cell parameters (which deviates from experiment). Duplicated ICSD entries, may skew representation toward certain crystal structures, but removing them would introduce a different bias by favoring one experimental verification over another, which is the reason for not excluding them.

3.2 MODEL

The main objective of the VAE is to learn the distribution of disordered inorganic crystals from the dataset, to eventually enable CSP. This procedure is illustrated in Figure 3. Firstly, the VAE needs to encode the representation from Figure 1 into the latent space. This is done through a convolution neural network (CNN) with three convolution layers for the atomic representation, and an Multilayer Perceptron (MLP) with two linear layers for the crystal representation. The output of the two networks are combined into $Z_{mean} \in \mathbb{R}^n$ and $Z_{var} \in \mathbb{R}^n$, which parameterize the multivariate Gaussian distribution in the latent space. Secondly, the decoder is trained to generate samples from the latent space and reconstruct them into the structural representation. This is done through another CNN for the atomic representation and an MLP for the crystal representation. At the end layer of the CNN, the output is divided into the parameters, partial occupation of atoms, Wyckoff multiplicity, disordered site indicator, fractional coordinates and Wyckoff letter, all of which have

their own loss function. At the end layer of the MLP for the crystal representation, the output is divided into the lattice parameters and the space group, which similarly have their own loss function.

In total, 7 loss functions are used for the reconstruction of the inorganic crystal, representing the 7 properties constructed during decoding (lattice parameters, space group, disordered site indicator, Wyckoff letter, Wyckoff multiplier, fractional coordinates and partial occupancy at each atomic site). This loss along with a Kullback-Leiber (KL) divergence loss $\mathcal{L}_{KL}$ constructs the total reconstruction loss. The total reconstruction loss encourages the model to generate the output data as close as possible to the input data. Detailed description of the loss function are presented in Appendix A

To illustrate the capability of Dis-CSP, the filtered ICSD, represented as in Figure 1, is used to train a VAE model. The dataset (138,692 structures) is randomly split into a validation set (10%) and test set (20%). The specifics of the VAE model can be considered in Appendix A.

4 GENERATION OF DISORDERED INORGANIC CRYSTALS

4.1 RECONSTRUCTION ERROR

From the test set, we assess the reconstruction errors associated with the encoding and decoding of Dis-CSP. As shown in Table 1, the reconstruction error of the lattice parameters, in mean absolute error (MAE), is small, as are the errors related to the space group and the disordered site indicator. Regarding the Wyckoff site parameters presented in Table 2, we achieve high accuracy for both the Wyckoff letter and Wyckoff multiplicity, for both the ordered and disordered Wyckoff sites. Similarly, the reconstruction error of the fractional coordinates remains small. For the partial occupancy, we use two different error metrics for the ordered and disordered cases. For the ordered structures, we can report an accuracy score, as a single type of atom occupies any Wyckoff site, making it a classification problem (correctly or wrongly occupied). However, for the disordered structures, several atoms can occupy the same Wyckoff sites, making it a regression problem and hence we report an MAE. We achieve a satisfactory result considering that the majority of inorganic structures exhibit occupancies greater than 5% , as illustrated in Figure 2. In future training of the model, a higher accuracy will be a valuable improvement, as experimental studies have demonstrated that even incorporating low concentrations of elements can enhance the functional properties of materialsAhaliabadeh et al. (2022); Chen et al. (2019b). Remarkably, the model does not overfit to either disordered or ordered Wyckoff sites, despite the bias present in the dataset, as shown in Figure 2. This is a significant achievement, as it demonstrates that the model is not biased towards any specific group of Wyckoff sites. Instead, it maintains a balanced representation, allowing for accurate predictions across the entire range of Wyckoff site configurations.

The Wyckoff site is determined by the site-symmetry and space group. Consequently, for a given Wyckoff letter and space group, the corresponding Wyckoff multiplicity is uniquely defined. Leveraging this, we evaluate the representations throughout the VAE by comparing the predicted Wyckoff multiplicity with the reference values determined by the space group and Wyckoff letter. The symmetry-matching accuracy (SMA) was found for the test set to be 99.5% and remains similar to the reconstructed accuracy of the Wyckoff multiplicity value in Table 2, implying the symmetry is preserved throughout the VAE.

A visualized representation of the reconstruction error can be considered in Appendix B.

Table 1: Reconstruction error of the test set for the lattice parameters (MAE), space group (Accuracy) and disordered site indicator (Accuracy).

Parameters	**Dis-CSP**
(a,b,c) [Å]	(0.06, 0.05, 0.10)
(α,β,γ) [°]	(0.02, 0.05, 0.28)
Space group	99.6%
Disordered site	99.8%

Table 2: Reconstruction errors of the test set for the parameters directly related to the disordered and ordered Wyckoff sites. The partial occupancy is presented with a MAE for the disordered sites, and an accuracy for the ordered sites. Accuracy is also used for the Wyckoff letter and Wyckoff multiplicity, while MAE is used for the fractional coordinate.

Parameters	Disordered Wyckoff sites	Ordered Wyckoff sites
Partial occupancy	0.05 (MAE)	99.1%
Wyckoff multiplicity	99.6%	99.8%
Wyckoff letter	99.3%	99.6%
Frac. coordinate [Å]	0.07	0.08

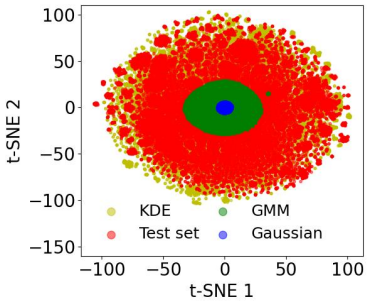


Figure 4: The latent space of the test set compared to a multivariate Gaussian sampling, along with a Kernel Density Estimation (KDE) Chen (2017) sampling and a Gaussian Mixture Model (GMM) Reynolds et al. (2009) sampling, both of which are trained on the latent space of the training set. The sample set for all models are of the same size as the test set.

4.2 Crystal structure prediction

Crystal structure prediction inherently carries the risk of generating physically meaningless structures. However, our representation is grounded in both crystal and site-symmetry, which allows us to minimize this issue during the reconstruction process. Specifically, by leveraging symmetry constraints, we can effectively avoid the formation of unreasonable inorganic crystal structures. The reconstruction process begins by generating the crystal cell from the predicted lattice parameters, thereby establishing the structural framework. Next, fractional coordinates and partially occupied atoms are incorporated, where the latter are assigned based on the partial occupancy if the disordered site indicator does not classify them as ordered sites. Once the structural framework and occupancy is established, we enforce symmetry constraints based on the predicted space group, Wyckoff letters, and Wyckoff multiplicity. This is done using Symmertrized structure group along with its symmetrical operations in Pymatgen Ong et al. (2013). If the reconstructed structure fails to satisfy these symmetry constraints, it is discarded, and the next candidate is considered.

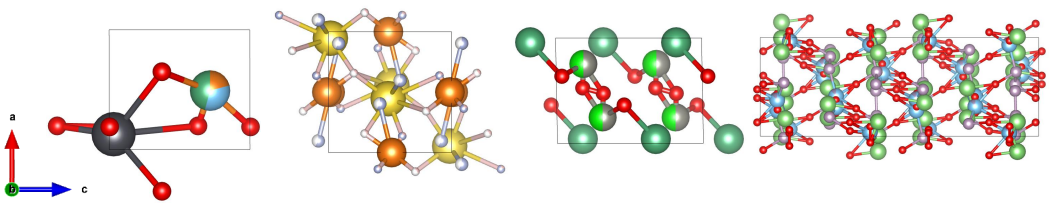


Figure 5: Four generated disordered inorganic crystal structures, using Dis-CSP. All crystals are viewed along the b-axis of the crystal.

Table 3: The generation error, defined as the percentage of sampled structures discarded during the reconstruction process, the symmetry matching accuracy (SMA), between the reconstructed and the symmetry required Wyckoff multiplicity for the three latent space estimators, and the validity, as the percentage of generated structures containing atomic sites closer than 0.5Å, are evaluated for the three latent space estimators using the test set as a reference.

Latent space	**Generation error**	**SMA**	**validity**
KDE	0%	99.6%	15.7%
GMM	2.64%	47.5%	16.4%
Gaussian	9.0%	23.0%	14-0%
Test set	0%	99.5%	14.0%

To perform CSP from the latent space of Dis-CSP, it is vital to accurately characterize its posterior distribution. The KL divergence loss encourages the formation of a smooth multivariate Gaussian distribution; however, this comes at the cost of an increased reconstruction loss. Given the complexity of representing disordered crystals, achieving a perfectly Gaussian latent space is inherently challenging. To further analyze the latent space, we use t-SNEVan der Maaten & Hinton (2008) to reduce the dimensionality of the latent space to two dimensions, facilitating visual comparisons of the test set latent space. To sample beyond the train set and test set, it is crucial to estimate parts of the latent space using different estimators. Beyond the conventional Multiveriate Gaussian sample, the Kernel Density Estimation (KDE) Chen (2017) and the Gaussian Mixture Model (GMM) Reynolds et al. (2009) are trained on the training datset to etsimate the overall latent space. These are vizualized along with the latent space in Figure 4. Visually, the KDE model offers the most accurate approximation of the latent space, providing a reasonable representation of its diverse distribution, whereas the GMM Reynolds et al. (2009) and the multivariate Gaussian distribution show less satisfactory results. This observation is further supported in Table 3, where the generation error, defined as the percentage of sampled structures discarded during reconstruction process, highlights differences between the three models.

While the reconstruction process eliminates most symmetrically invalid structures, it does not fully guarantee that the reconstructed structure obey the exact symmetries of the crystal. Specifically, discrepancies may arise where the predicted Wyckoff multiplicity differs from the symmetry-constrained values determined by the space group and the Wyckoff letter. To evaluate this, we analyze the SMA presented in Table 3. These results clearly indicate that sampling from the KDE-estimated latent space ensure a symmetrical equivariance. Another important aspect to consider during the reconstruction process is that symmetrical sites must be positioned based on their fractional coordinates. This can introduce errors when two sites are too close to each other. The position error (PE), as presented in Table 3, clearly indicates that such filtering is necessory to the latent space estimators. All of this underscores the superior performance of KDE in capturing the complexity of the latent space, while maintaining equivarient representation.

Using the KDE estimation of the latent space, we can generate disordered structures, with four representative generations illustrated in Figure 5. Given the stochastic nature of our sampling process, the generated structures appear reasonable. The leftmost crystal, $Mg_{0.38}Ti_{0.64}Nb_{0.88}Pb_2O_6$, suggests a doping strategy involving Mg, Ti, and Nb, elements commonly used to enhance functional materials for catalysis and biomedical applications Cui et al. (2014); Li et al. (2021). The generated $Na_{7.92}Mg_8H_{3.76}F_{10.88}$ and $Li_{36}Ti_{30.96}Sn_{4.68}P_{36}O_{72}$ structures resemble electrode materials for batteries. The latter, in particular, closely aligns with known lithium-ion battery anode or cathode materials, featuring doping at the metallic (Ti) site, although Sn may not be an appropriate dopant Weng et al. (2017). Additionally, the generated $Zr_2Nb_{3.96}Zn_2O_8$ structure is structurally related to the $ZrNbO_4$ alloy Peyret et al. (2023), with Zn partially occupying the Zr site, suggesting a potential doping strategy for this material. Despite the promising nature of the generated structures, certain compositions, such as $Na_{7.92}Mg_8H_{3.76}F_{10.88}$, appear chemically unreasonable. This indicates the necessity of incoporating chemically intuitive filtering into the CSP process to eliminate unrealistic structures. Currently, no established chemical intuition filters exist for disordered crystals. However, we are actively developing such methods, which will assess the likelihood of partial occu-

pations of atoms at symmetrical sites based on factors such as atomic charge and chemical bonding characteristics.

A limitation of using a VAE model for CSP is that the latent space is constrained by the distribution of the training set. While KDE is highly effective at interpolating within known distributions, it struggles with extrapolation, which could be addressed by using diffusion models or transformers. This explains why the generated structures in Figure 5 closely resemble battery electrodes and alloys, since they are among the most investigated materials, particularly in the context of disordered inorganic crystals.

Despite this, the ICSD dataset encompasses a vast and diverse range of experimentally verified inorganic crystals. This extensive coverage suggests that numerous potential crystal structures remain unexplored in the ICSD distribution, presenting significant opportunities for the discovery of novel functional materials. This is particularly true for disordered crystals, where possible doping configurations are virtually limitless. However, blindly searching through this vast chemical space is inefficient. To guide the search towards promising candidates, a property-oriented approach incorporating physical and chemical constraints is necessary.

However, computing physical or chemical properties for disordered materials has its own challenges. Finite structure sizes are needed to computationally calculate the properties, but crystal symmetry breaks down when assigning partial occupancy to the symmetrical sites. To create a property distribution and compare it to experimental data, one needs to sample enough configurations of the atoms at large enough supercell sizes; and even then, the comparison may be off as shown in Appendix C. Such property distributions, dependent on a large sample size of the disordered crystals, are rarely accounted for in theoretical databases like the Materials Project database **?**. Even using machine learning force fields for property predictions are not feasible and external methods are needed to accurately map the properties Xie et al. (2024). For these reasons, property-oriented searches have been unexplored for for disordered inorganic crystal structures.

In our approach, the latent space of the VAE model allows an opportunity to bypass this issue and condition our search on specific compositions, space groups or even target certain partial occupations of atoms. By defining specific terms for the desired configurations, we employ gradient descent to optimize the latent sampling z, with respect to the target configuration, as illustrated in Appendix D . This is feasible because Dis-CSP reconstructs multiple parameters related to both crystal symmetry and atomic representation.

5 Conclusion

In this work we introduce **Dis-CSP: Disordered crystal structure prediction**, the world's first generative model designed for disordered inorganic crystal structures, leveraging a novel representation for disordered crystals, incorporating partial occupancy at symmetrical sites. To demonstrate the capabilities of Dis-CSP, we utilize the Inorganic Crystal Structure Database (ICSD) - the largest collection of experimentally inorganic crystals - to train a generative model, enabling crystal structure prediction for disordered crystals. Due to the equivariant representation of crystal structures, Dis-CSP inherently filters out symmetry-violating structures, ensuring a symmetry-consistent approach to generating valid structures. Moreover, Dis-CSP offers a promising alternative framework to the practically-infeasible property-oriented search, instead targeting compositional similarity for latent space exploration of disordered inorganic structures. With Dis-CSP, we initiate the systematic exploration of disordered inorganic crystals, aiming to discover novel structures with potential applications across various scientific fields.

References

A Abdelghany, SN Elsayed, DM Abdelwahab, AH Abou El Ela, and NH Mousa. Electrical conductivity and thermoelectric power of agsbse2 in the solid and liquid states. *Materials chemistry and physics*, 44(3):277–280, 1996.

Zahra Ahaliabadeh, Xiangze Kong, Ekaterina Fedorovskaya, and Tanja Kallio. Extensive comparison of doping and coating strategies for ni-rich positive electrode materials. *Journal of Power Sources*, 540:231633, 2022.

Mila AI4Science, Alex Hernandez-Garcia, Alexandre Duval, Alexandra Volokhova, Yoshua Bengio, Divya Sharma, Pierre Luc Carrier, Yasmine Benabed, Michał Koziarski, and Victor Schmidt. Crystal-gfn: sampling crystals with desirable properties and constraints. *arXiv preprint arXiv:2310.04925*, 2023.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):1–16, 2024.

Jin Hyun Chang, David Kleiven, Marko Melander, Jaakko Akola, Juan Maria Garcia-Lastra, and Tejs Vegge. Clease: a versatile and user-friendly implementation of cluster expansion method. *Journal of Physics: Condensed Matter*, 31(32):325901, 2019.

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, 2019a.

Jie Chen, Huiping Yang, Tianhao Li, Chaoyang Liu, Hui Tong, Jiaxin Chen, Zengsheng Liu, Lingfeng Xia, Zhaoyong Chen, Junfei Duan, et al. The effects of reversibility of h2-h3 phase transition on ni-rich layered oxide cathode for high-energy lithium-ion batteries. *Frontiers in Chemistry*, 7:500, 2019b.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

Kaibin Chu, Wei Zong, Guohao Xue, Hele Guo, Jingjing Qin, Haiyan Zhu, Nan Zhang, Zhihong Tian, Hongliang Dong, Yue-E Miao, et al. Cation substitution strategy for developing perovskite oxide with rich oxygen vacancy-mediated charge redistribution enables highly efficient nitrate electroreduction to ammonia. *Journal of the American Chemical Society*, 145(39):21387–21396, 2023.

François Cornet, Federico Bergamin, Arghya Bhowmik, Juan Maria Garcia Lastra, Jes Frellsen, and Mikkel N. Schmidt. Kinetic langevin diffusion for crystalline materials generation. In *AI for Accelerated Materials Design - ICLR 2025*, 2025. URL https://openreview.net/forum?id=Mttf1RoKKM.

Jie Cui, Jiangwen Liu, Hui Wang, Liuzhang Ouyang, Dalin Sun, Min Zhu, and Xiangdong Yao. Mg–tm (tm: Ti, nb, v, co, mo or ni) core–shell like nanostructures: synthesis, hydrogen storage performance and catalytic mechanism. *Journal of Materials Chemistry A*, 2(25):9645–9655, 2014.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

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

K Funke, RD Banhatti, Paweł Grabowski, J Nowinski, Wojciech Wróbel, R Dinnebier, and O Magdysyuk. Low-temperature α-agi confined in glass: Structure and dynamics. *Solid State Ionics*, 271:2–9, 2015.

Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, 2004.

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*, 36:17464–17497, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *ICLR*, 2024.

Nikita Kazeev, Ruiming Zhu, Ignat Romanov, Andrey E Ustyuzhanin, Shuya Yamazaki, Wei Nong, and Kedar Hippalgaonkar. Wyckofftransformer: Generation of symmetric crystals. In *AI for Accelerated Materials Design-NeurIPS 2024*.

Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alan Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *ACS central science*, 6(8):1412–1420, 2020.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Zhongjie Li, Hao Xu, Anping Dong, Jiajun Qiu, Lin He, Ting Zhang, Dafan Du, Hui Xing, Guoliang Zhu, Donghong Wang, et al. Characteristics of ti-nb-mg alloy by powder metallurgy for biomedical applications. *Materials Characterization*, 173:110953, 2021.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. In *Forty-first International Conference on Machine Learning*, 2024.

Cai Ning, Qun Ji, Yilei Wu, Jinlan Wang, and Ming-Gang Ju. Disorder on mixed cation halide perovskite for photovoltaic applications. *The Journal of Physical Chemistry Letters*, 14(36):8034–8042, 2023.

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.

Duncan Peyret, Damien Kaczorowski, Milan Skocic, Bernard Tribollet, and Vincent Vivier. Electrochemical and modelling study of zrnbo alloys aged under high temperature and high pressure pwr simulated conditions. *Corrosion Science*, 224:111505, 2023.

Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

Steven H Simon. *The Oxford solid state basics*. OUP Oxford, 2013.

Andrew E Smith, Hiroshi Mizoguchi, Kris Delaney, Nicola A Spaldin, Arthur W Sleight, and Mas A Subramanian. Mn3+ in trigonal bipyramidal coordination: a new blue chromophore. *Journal of the American Chemical Society*, 131(47):17084–17086, 2009.

B Souvignier. General and special wyckoff positions. 2015.

Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

Tianyu Su, Brian J Blankenau, Namhoon Kim, Jessica A Krogstad, and Elif Ertekin. First-principles and cluster expansion study of the effect of magnetism on short-range order in fe-ni-cr austenitic stainless steels. *Acta Materialia*, pp. 120088, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Heng Wang, Aaron D LaLonde, Yanzhong Pei, and G Jeffery Snyder. The criteria for beneficial disorder in thermoelectric solid solutions. *Advanced Functional Materials*, 23(12):1586–1596, 2013.

Zhaoyang Wang, Zijuan Du, Luoqing Wang, Guanjie He, Ivan P Parkin, Yanfei Zhang, and Yuanzheng Yue. Disordered materials for high-performance lithium-ion batteries: A review. *Nano Energy*, pp. 109250, 2024.

Guo-Ming Weng, Long-Yin Simon Tam, and Yi-Chun Lu. High-performance liti 2 (po 4) 3 anodes for high-areal-capacity flexible aqueous lithium-ion batteries. *Journal of Materials Chemistry A*, 5(23):11764–11771, 2017.

Hsin-jay Wu, Sinn-wen Chen, Teruyuki Ikeda, and G Jeffrey Snyder. Reduced thermal conductivity in pb-alloyed agsbte2 thermoelectric materials. *Acta Materialia*, 60(17):6144–6151, 2012.

Jun-Zhong Xie, Xu-Yuan Zhou, Bin Jin, and Hong Jiang. Machine learning force field-aided cluster expansion approach to phase diagram of alloyed materials. *Journal of Chemical Theory and Computation*, 20(14):6207–6217, 2024.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *ICLR*, 2022.

Adrian Xiao Bin Yong, Tianyu Su, and Elif Ertekin. Dismai-bench: benchmarking and designing generative models using disordered materials and interfaces. *Digital Discovery*, 3(9):1889–1909, 2024.

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

Peichen Zhong, Sunny Gupta, Bowen Deng, KyuJung Jun, and Gerbrand Ceder. Effect of cation disorder on lithium transport in halide superionic conductors. *ACS Energy Letters*, 9:2775–2781, 2024.

Ruiming Zhu, Siyu Isaac Parker Tian, Zekun Ren, Jiali Li, Tonio Buonassisi, and Kedar Hippalgaonkar. Predicting synthesizability using machine learning on databases of existing inorganic materials. *ACS omega*, 8(9):8210–8218, 2023.

Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. Wycryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488, 2024.

A VAE MODEL SPECIFICS

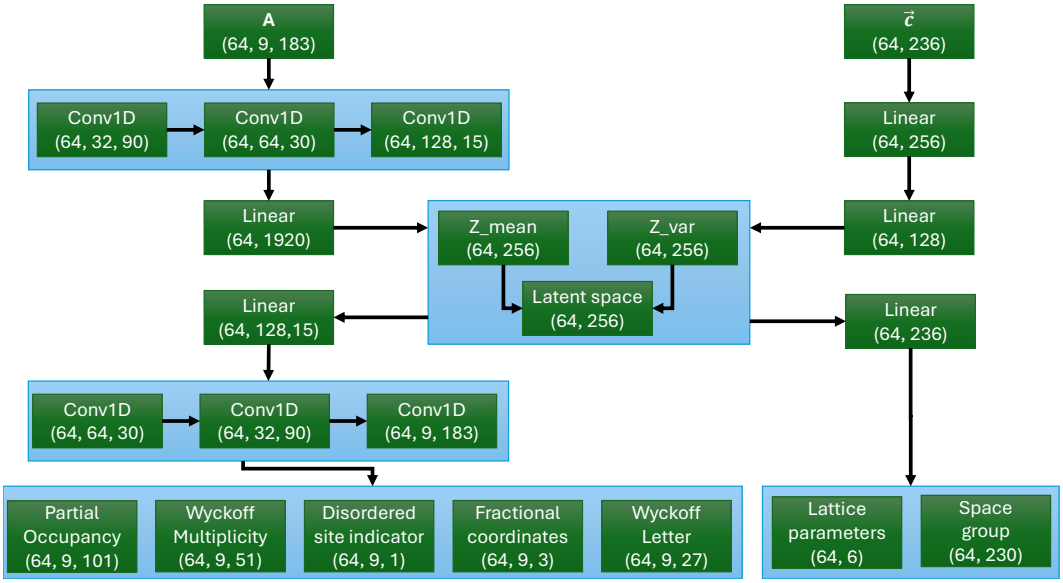


Figure 6: The architecture of the VAE model used for Dis-CSP with a batch size of 64. In the schematic representation, green blocks denote individual layers within the VAE, while blue blocks represent groups of layers. The model accepts two inputs: (1) matrix $\boldsymbol{A}$, which encodes the atomic representation and (2) a vector $\boldsymbol{c}$, which encodes the crystal representation. These inputs are processed separately by the encoder, then concatenated in the latent space and then separated in the decoder. The atomic representation produces five outputs, while the crystal representation yields two two outputs. The Relu activation function is utilized for the CNN, while the sigmoid activation function is utilized for the disordered site indicator and the softmax activation function is utilized for the partial occupancy, Wyckoff multiplicity, Wyckoff letter and Space group.

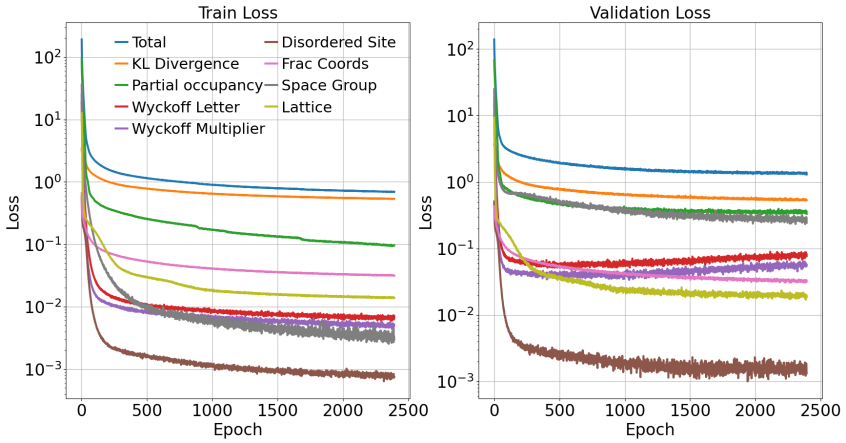


Figure 7: The variational auto-encoder (VAE) training curves, with curves color-coded uniformly across the plots.

The total reconstruction loss function consists of 7 loss functions $\mathcal{L}_{recon}$, along with a Kullback-Leiber (KL) divergence loss $\mathcal{L}_{KL}$.

For the reconstruction loss, two different loss functions are used:

Table 4: Loss coefficients used for training the VAE model used in this paper.

Coefficients	Optimal value
λ_{KL}	1.0
λ_{spg}	10.0
$\lambda_{lattice}$	3.0
λ_{occ}	2000
λ_{mult}	1.0
λ_{letter}	1.0
$\lambda_{disorder}$	0.1
λ_{coord}	1.0

The cross entropy loss function, which calculates the likelihood of getting the reconstructed distribution $\hat{Q}$, given the input distribution Q:

$$\mathcal{H}(\hat{Q}|Q) = -\sum_i Q(i)\log(\hat{Q}(i)) \tag{2}$$

The mean square error (MSE) loss function compares the difference between the input values Y and the reconstructed values $\hat{Y}$:

$$MSE(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2 \tag{3}$$

For the parameters in the atomic representation, the loss functions are taken per Wyckoff site and the total loss is a sum of all individual losses. The Wyckoff letter loss $\mathcal{L}_{letter}$, the Wyckoff multiplicity loss $\mathcal{L}_{mult}$ and disordered site indicator loss $\mathcal{L}_{disorder}$ use the cross entropy loss function, while the fractional coordinates loss $\mathcal{L}_{coord}$ and the partial occupancy loss $\mathcal{L}_{occ}$ use the MSE loss function. Note that for partial occupancy, the loss is defined by the difference between the distributions of atoms at a Wyckoff site.

For the parameters in the crystal representation, the loss functions are used directly on the representation. The lattice parameter loss $\mathcal{L}_{lattice}$ uses the MAE loss function, while the space group loss $\mathcal{L}_{spg}$ uses the cross entropy loss function.

The KL divergence loss $\mathcal{L}_{KL}$ is used to shape the latent space into a Gaussian distribution. The KL divergence loss calculates the difference between $q(z|X)$, the learned distribution of latent points z given input data X, and $p(z)$, the desired Gaussian distribution for the latent points, Kingma (2013); Kullback & Leibler (1951):

$$\mathcal{L}_{KL} = KL(q(z|X)||p(z)) \tag{4}$$

During the optimization step of the VAE model the total loss function

$$\begin{aligned}\mathcal{L} =& \lambda_{KL}\mathcal{L}_{KL} + \lambda_{spg}\mathcal{L}_{spg} + \lambda_{lattice}\mathcal{L}_{lattice} + \lambda_{occ}\mathcal{L}_{occ} \\ &+ \lambda_{mult}\mathcal{L}_{mult} + \lambda_{letter}\mathcal{L}_{letter} + \lambda_{disorder}\mathcal{L}_{disorder} + \lambda_{coord}\mathcal{L}_{coord}\end{aligned} \tag{5}$$

is optimized, with λ_i as the coefficient for each loss contribution.

Figure 6 illustrates the architecture of the VAE model used for generating disordered inorganic crystals. The VAE takes two inputs: (1) a matrix $\boldsymbol{A}$ describing the atomic representation and (2) a vector $\boldsymbol{c}$ describing the crystal representation. The model produce seven output: partial occupancy, Wyckoff multiplicity, disordered site indicator, fractional coordinates, Wyckoff letter, lattice parameters and space group. To optimize the model architecture, hyperparameter tuning was conducted for the CNN to determine the optimal number of layers, channel dimensions and signal dimensions. Similarly, for the MLP, hyperparameter tuning was performed to optimize the number of hidden layers and nodes.

For the training, the coefficients in the total loss function. in Equation (5), were optimized to ensure robust performance on the test set while preventing overfitting, as assessed by validation loss. The final coefficients, presented in Table 4, were determined to provide the most effective balance. Notably, partial occupancy was the most challenging parameter to optimize, resulting in a higher loss

coefficient to improve accuracy. In contrast the difference between the other coefficients are relative small. The disordered site was the easiest parameter to predict, which justify assigning it the lowest coefficient.

Training was conducted using Adam optimizer Kingma (2014) with a learning rate of 5×10^{-6} and a batch size of 64. The model was trained fro 2500 epochs, with the final VAE model selected based on lowest validation loss, which occurred at epoch 2349.

Figure 7 visualizes the training process, where it is evident that the KL loss dominates among the eight loss function. This behavior is anticipated, as the KL loss must balance the combined effect of all reconstruction-related loss term. Some variations are observed in specific loss functions, which could potentially be avoided by using a lower learning rate. However, these variations are minimal and can be considered negligible.

B VISUALIZATION OF THE RECONSTRUCTION ERROR

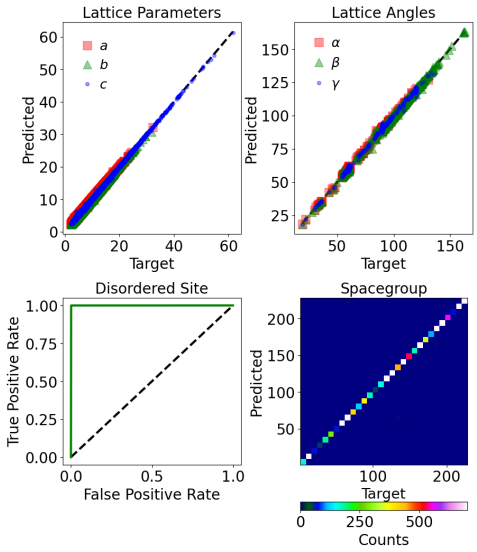


Figure 8: The lattice parameter prediction along with the disordered site prediction. Note the perfect ROC curve stems from the fact that only 1% of the Wyckoff sites were misclassified.

By plotting the reconstructed representation against the target representation for the test set, obvious outliers can be identified, providing insight into the quantitative results presented in Table 1 and Table 2.

In Figure 8, the reconstructed lattice parameters, space group and disordered site indicator are compared to their target values. No obvious outliers are detected in the lattice parameters or space group, and the ROC curve for the disordered site indicator does not indicate significant errors.

In Figure 9, the partial occupancy, Wyckoff letter, Wyckoff multiplicity and fractional coordinates are compared to the target values, with a distinction between disordered and ordered Wyckoff sites. For the Wyckoff letter and Wyckoff multiplicity, no obvious outliers are detected. The fractional coordinates exhibit noise around the 1:1 line, consistent with the error rate in Table 2, and this noise does not differ significantly between the disordered and ordered Wyckoff sites. The partial occupancy also displays noise around the 1:1 line in both cases, though it is more pronounced for the disordered Wyckoff sites. This observation aligns with the quantitative results in Table 2 and suggests that higher accuracy may be required in future training of the VAE model. However, achieving this improvement is challenging due to the high diversity in partial occupancy per Wyckoff site within the dataset. To enhance accuracy, additional filtering strategies may be necessary to refine the dataset, or other training strategies may be needed.

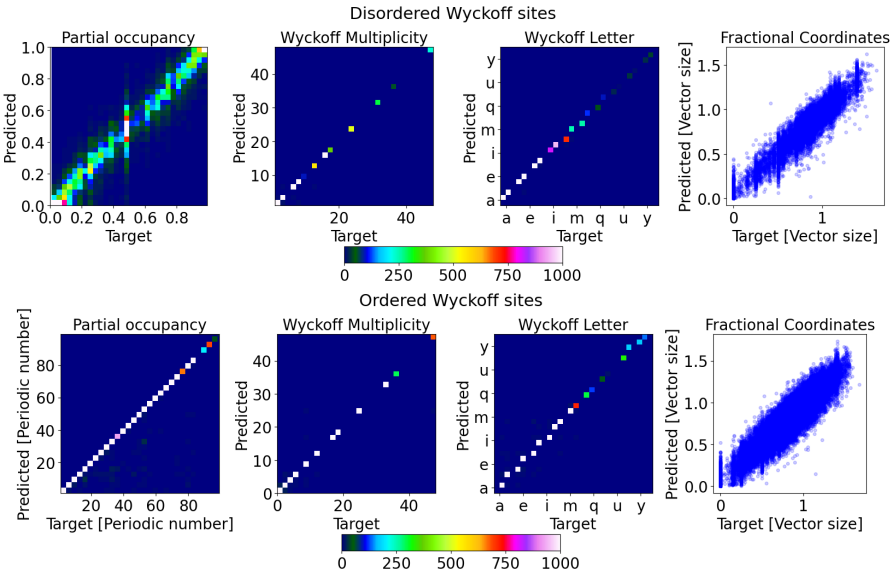


Figure 9: The ordered and disordered Wyckoff site predictions along the respected color bars. **Top**: The disordered Wyckoff sites, with partial occupation of atoms along with its other characteristics. **Bottom**: The ordered Wyckoff sites with an occupation of a single atom along with its other characteristics.

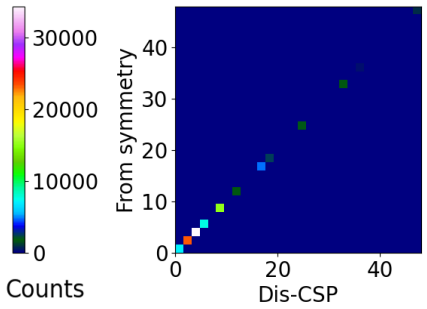


Figure 10: Comparison between the reconstructed and the symmetry required Wyckoff multiplicity, showing an symmetry matching accuracy (SMA) of 99.5%.

Moreover, the symmetry matching accuracy (SMA) can be visualized by plotting the recontructed Wyckoff mulitpliers, but the one defined from the reconstructed Wyckoff letter and space group as illustrated in Figure 10

C DISTRIBUTION OF PROPERTIES

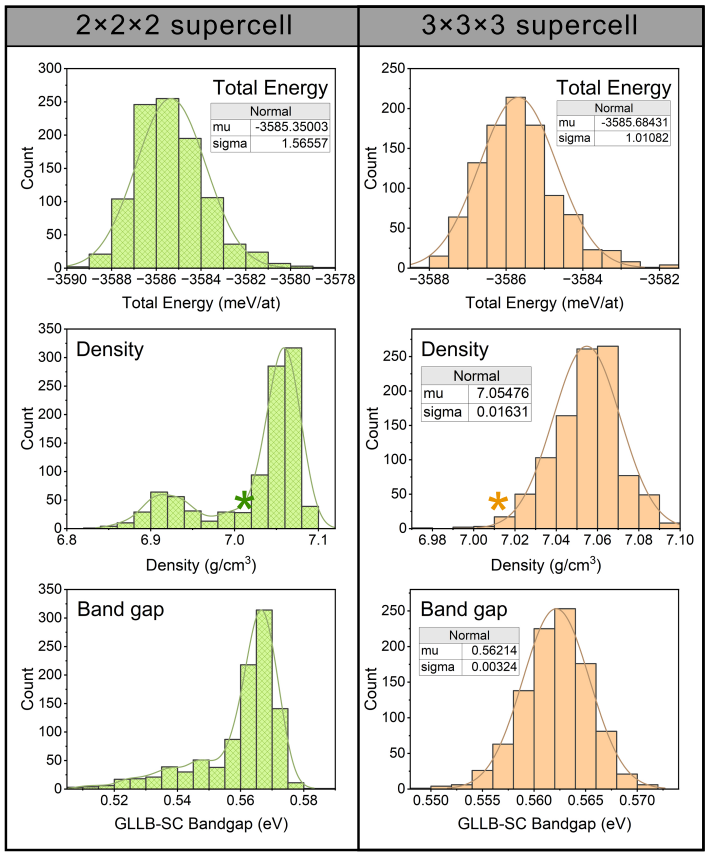
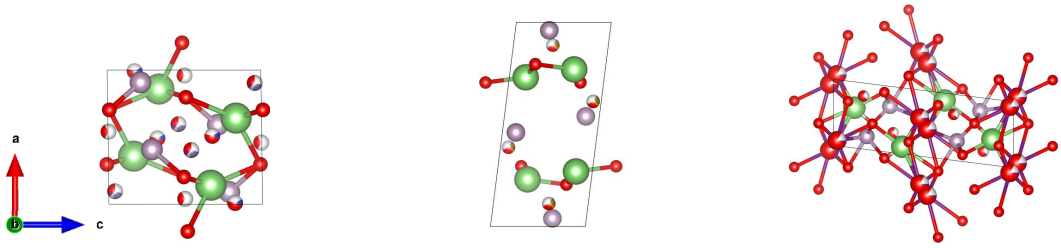


Figure 11: Statistical distributions of selected properties of 1000 virtual cells of site-disordered $AgSbTe_2$. The effect of supercell size (doubled vs. tripled) is compared. In cases where the distribution is approximately normal, the mean (mu) and standard deviation (sigma) are specified. For comparison, the experimental band gap of $AgSbTe_2$ is 0.71 eV Abdelghany et al. (1996) and the experimental density of $AgSbTe_2$ is 7.012 g/cm^3 (marked with asterisk) Wu et al. (2012).

Computed properties of disordered inorganic crystals with partial occupancy are not directly accessible by current first-principles methods or machine-learned force fields. Rather, we estimate these properties by generating a set of *virtual cells*, based on the crystal representation. A sufficiently large size of the virtual cell, measured in terms of supercell size, along with a sufficiently large number of virtual cells used to sample the configurational space, can approximate the physical properties of the disordered inorganic crystal. Specifically, given a temperature T and using Maxwell-Boltzmann statistics, we can recover the expectation value $\langle P \rangle$ of a certain property P from the calculated or predicted values p_i of each virtual cell (of energy E_i) in the sample set.

$$\langle P \rangle = \sum_i \frac{p_i e^{\frac{-E_i}{k_B T}}}{\sum_j e^{\frac{-E_j}{k_B T}}} \tag{6}$$

Ignoring the effects of temperature, the properties are better represented as distributions, rather than as single number figures. We give an example of the properties of disordered $AgSbTe_2$ (Figure 11) with Ag and Sb being partially occupied at the same Wyckoff site. Total energies and densities are calculated using CHGNET Deng et al. (2023), and the band gaps are predicted using the GLLB-SC model in MEGNet Chen et al. (2019a). We note that generating virtual cells from the doubled cell creates skewed or bimodal distributions of density and band gap deviating from those of the tripled supercell, indicating the role of interference from the periodic boundary.



$Li_4O_{17.12}P_8Cu_{1.04}Ge_{1.84}Rh_{0.72}I_{0.8}$
Space Group: 62

$Li_8P_8O_{10.24}Tl_{1.12}Fe_{0.88}Au_{0.48}F_{0.4}$
Space Group: 62

$Li_8P_8O_{27.68}I_{0.48}F_{0.48}$
Space Group: 62

Figure 12: Three generated disordered inorganic crystal structures with the space group 62, 1 disordered Wyckoff site and the elements Li, P and O. All crystals are viewed along the b-axis of the crystal.

D CONDITIONED CRYSTAL STRUCTURE GENERATION

It is possible to condition crystal structure generation based on the framework of Dis-CSP. So in reality it is possible to choose any target value within our representation of the crystal structure. For example, we utilize gradient descent to explore the latent space, enabling the generation of structures with a space group of 62, a single disordered site, and with the elements Li, P, and O positioned at one of the Wyckoff sites. This is achieved by defining representing and defining loss function for each condition, and the Adam optimizer to optimize the latent space according to the specified criteria. Generation based on area of the latent space following this condition is illustrated in Figure 12 . If we want to further condition, we use the representation and the loss described in Section 3.2.