Space Group Informed Transformer for Crystalline Materials Generation

Zhendong Cao,[1,2] Xiaoshan Luo,[3,4] Jian Lv,[3,*] and Lei Wang[1,5,†]

[1]*Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[2]*School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China*
[3]*Key Laboratory of Material Simulation Methods and Software of Ministry of Education, College of Physics, Jilin University, Changchun 130012, P. R. China*
[4]*State Key Laboratory of Superhard Materials, College of Physics, Jilin University, Changchun 130012, P. R. China*
[5]*Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China*

(Dated: March 26, 2024)

We introduce `CrystalFormer`, a transformer-based autoregressive model specifically designed for space group-controlled generation of crystalline materials. The space group symmetry significantly simplifies the crystal space, which is crucial for data and compute efficient generative modeling of crystalline materials. Leveraging the prominent discrete and sequential nature of the Wyckoff positions, `CrystalFormer` learns to generate crystals by directly predicting the species and locations of symmetry-inequivalent atoms in the unit cell. Our results demonstrate that `CrystalFormer` matches state-of-the-art performance on standard benchmarks for both validity, novelty, and stability of the generated crystalline materials. Our analysis also shows that `CrystalFormer` ingests sensible solid-state chemistry information from data for generative modeling. The `CrystalFormer` unifies symmetry-based structure search and generative pre-training in the realm of crystalline materials. The simplicity, generality, and flexibility of `CrystalFormer` position it as a promising architecture to be the foundational model of the entire crystalline materials space, heralding a new era in materials modeling and discovery.

I. INTRODUCTION

Machine learning methods are playing an increasingly important role in material discovery, complementing conventional computational approaches [1, 2]. Generative machine learning, in particular, has been a promising step for matter inverse design [3, 4] which goes beyond machine learning accelerated structure search [5] and property screening [6]. Generative models learn the underlying distribution of training data and generated new samples from the learned distribution. In addition, the generation process can also be controlled by prompts such as desired material properties or experiment observations. Amazing programming abilities of generative models have been demonstrated in large language model [7], text-to-image generation [8, 9], and protein design [10].

It is anticipated that generative model-based approaches will introduce groundbreaking changes to the traditional workflows of material discovery. A generative pretrained foundation model for crystalline materials is a key step towards such lofty goal. However, despite of intensive efforts [11–22], the current generative models for crystalline materials fall short to match the success of other domains. Simply scaling the compute and model size of the current crystal generative model may not be enough because the amount of high-quality data for crystalline materials is much less than compared to language and image domains. Therefore, leveraging the inherent inductive biases specific to crystalline structures for more data-efficient generative modeling is essential, as has been pursued in some of recent works [12, 23–26].

$P1$ world	With space group symmetry
$(100 \times 100^3)^{20} \approx 10^{160}$	$(100 \times 10 \times 100)^5 \approx 10^{25}$

TABLE I. A back-of-envelope estimate of the size of the crystalline material space. In the "$P1$ world", one treats crystals as if they were in the least symmetric $P1$ space group. We consider 100 possible chemical elements and 20 atoms in the unit cell with 100 discretizations in each direction. While in the case of utilizing the symmetry of a typical space group, we consider 5 symmetry inequivalent atoms occupying 10 possible Wyckoff positions. The additional factor of 100 account for the remaining degree of freedoms for the fractional coordinates. See Refs. [27, 28] for alternate estimates of the materials space in the context of crystal structure prediction.

The space group symmetry is arguably the most important inductive bias in the modeling of crystalline materials, which is the joint outcome of the rotational and translational symmetry in space. There are in total 230 space groups [29] for three-dimensional crystal structures. Nature exhibits a preference for symmetric crystal structures, a tendency that may be attributed to the symmetry inherent in the interatomic interactions, which, in turn, are governed by the fundamental forces acting between elementary particles. As a result, the appearance of crystalline materials in the the first and the least symmetric space group $P1$ is rare [30], with many instances potentially being misclassified. [31].

Space group symmetry imposes significant constraints on a crystal. At its core, the space group identifies the crystal system to which a crystal belongs, thereby limiting the permissible values for the lattice parameters that define the dimensions and angles of the crystal's unit cell. Moreover, the symmetry operations associated with a given space group ensure that identical atoms are consistently mapped onto each other

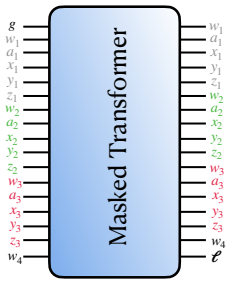
* lvjian@jlu.edu.cn
† wanglei@iphy.ac.cn

(a)

Wyckoff Positions of Group R-3c (No. 167) [hexagonal axes]

Multiplicity	Wyckoff letter	Fractional coordinates $(0,0,0)+(\frac{2}{3},\frac{1}{3},\frac{1}{3})+(\frac{1}{3},\frac{2}{3},\frac{2}{3})+$
36	f	$(x,y,z),(-y,x-y,z),$ $(-x+y,-x,z),(y,x,-z+\frac{1}{2}),$ $(x-y,-y,-z+\frac{1}{2}),(-x,-x+y,-z+\frac{1}{2}),$ $(-x,-y,-z),(y,-x+y,-z),$ $(x-y,x,-z),(-y,-x,z+\frac{1}{2}),$ $(-x+y,y,z+\frac{1}{2}),(x,x-y,z+\frac{1}{2})$
18	e	$(x,0,\frac{1}{4}),(0,x,\frac{1}{4}),$ $(-x,-x,\frac{1}{4}),(-x,0,\frac{3}{4}),$ $(0,-x,\frac{3}{4}),(x,x,\frac{3}{4})$
18	d	$(\frac{1}{2},0,0),(0,\frac{1}{2},0),$ $(\frac{1}{2},\frac{1}{2},0),(0,\frac{1}{2},\frac{1}{2}),$ $(\frac{1}{2},0,\frac{1}{2}),(\frac{1}{2},\frac{1}{2},\frac{1}{2})$
12	c	$(0,0,z),(0,0,-z+\frac{1}{2}),$ $(0,0,-z),(0,0,z+\frac{1}{2})$
6	b	$(0,0,0),(0,0,\frac{1}{2})$
6	a	$(0,0,\frac{1}{4}),(0,0,\frac{3}{4})$

(b)



i	w_i	a_i	x_i	y_i	z_i
1	a	C	0	0	$\frac{1}{4}$
2	b	Ca	0	0	0
3	e	O	*0.257*	0	$\frac{1}{4}$
4	X	-	-	-	-

(c)

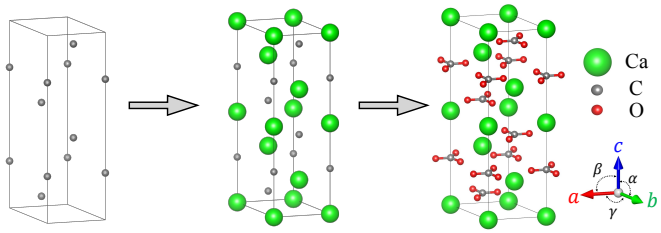


FIG. 1. (a) The Wyckoff positions of the $R\bar{3}c$ space group (No. 167). We highlight the occupied Wyckoff positions of calcite $CaCO_3$ crystal which belongs to this space group. Carbon, calcium, and oxygen atoms occupy the '6a', '6b', and '18e' positions, respectively. (b) The `CrystalFormer` and the essential crystal data of $CaCO_3$. The `CrystalFormer` is a decoder-only autoregressive transformer that models the space group controlled crystal structures by predicting probabilities of the Wyckoff letter w_i, chemical element a_i, and fractional coordinates (x_i, y_i, z_i) of each atom, and finally the lattice parameters $\boldsymbol{L}$ sequentially. In the table, $x_3 = 0.257$ is the only continuous variable that needs to be predicted. All other fractional coordinates are fixed by discrete data like the space group number and Wyckoff letter. (c) The autoregressive sampling procedure of the $CaCO_3$ crystal. One first places carbon atoms at the '6a' position, then places calcium atoms at '6b' position and finally places oxygen atoms at '18e' position. In each step of the sampling procedure, there is a choice of the Wyckoff positions, chemical element, and the fractional coordinates (if there are any remaining degrees of freedom in the Wyckoff position) of the atom.

across the crystal, preserving its structural integrity. This requirement enforces strict conditions regarding the types of chemical elements present, their specific locations within the crystal, and the total number of atoms in each unit cell. A key concept in understanding these constraints is the Wyckoff positions, which delineate unique areas within a unit cell that are defined by the symmetry operations of the crystal's space group. These positions are represented as fractional coordinates, enabling precise definition relative to the unit cell's axes. For example, Fig. 1(a) shows the Wyckoff positions for the space group $R\bar{3}c$ (No. 167). The Wyckoff positions are labeled by letters in the alphabet, starting from special points in the bottom to general positions in the top. The multiplicity counts the number of equivalent positions connected by the space group symmetry operations. All of them should be occupied by the same type of atoms to uphold the space group symmetry. For example, the top row of the table in Fig. 1(a) shows a general position (x, y, z) that can be mapped to 36 positions under the symmetry operations of the $R\bar{3}c$ space group.

Nature tends to place atoms in those special Wyckoff positions at the bottom of the table. For example, we highlight the occupied Wyckoff positions of calcite ($CaCO_3$) crystal in Fig. 1, associated with the $R\bar{3}c$ space group. One sees that the Wyckoff letter '6a' and '6b' deterministicly define the locations of the carbon and calcium atoms within the unit cell. In addition, it follows that $a = b$, and $\alpha = \beta = 90°, \gamma = 120°$ as the $R\bar{3}c$ space group belongs to the trigonal crystal system. Ultimately, despite having 30 atoms in the unit cell, there are only three continuous degrees of freedom for the $CaCO_3$ structure: the coordinate of oxygen atom $x = 0.257$ and the lattice constants $a = b = 4.99$Å and $c = 17.07$Å. All other information about the crystal structure can be specified via discrete data such as the Wyckoff letters and chemical species.

The prominent discrete and sequential features illustrated in Figure 1 are ubiquitous in crystalline materials. The Wyckoff positions not only specify possible locations of atoms in the unit cell, their associated multiplicities also put strong constraints on the number of atoms. Therefore, space group symmetry significantly reduces the degrees of freedom of crystalline materials. Failing to exploit this information in generative modeling not only renders learning inefficient, it also severely impairs the generalization ability of the model. For example, the performance of the generative model quickly deteriorates as the number of atoms increases due to it is challenging to generate highly symmetric crystal structures [16].

In this paper, we introduce `CrystalFormer`, an autoregressive transformer for generative modeling of crystalline materials. `CrystalFormer` models the joint probability distribution of Wyckoff positions, chemical species, and lattice parameters of crystals with a given space group. By treating the Wyckoff positions as the first class citizen in the model, `CrystalFormer` seamlessly integrates the space group symmetry into crystal probabilistic modeling. The space group-informed transformer exploits the fundamental inductive bias of crystals to greatly constrain and simplify the generation process. As shown in Table I, explicit modeling of the Wyck-

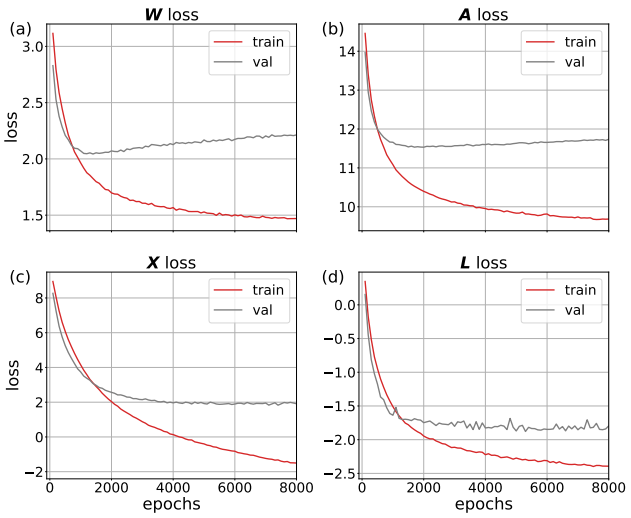


FIG. 2. Break up of the training and validation losses for (a) Wyckoff letters, (b) chemical species, (c) fractional coordinates, and (d) lattice parameters over training epochs.

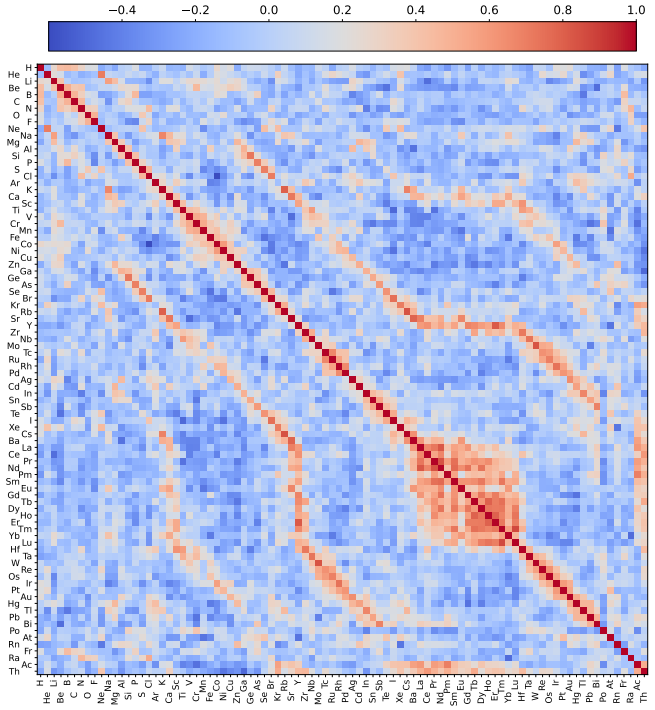


FIG. 3. The cosine similarity matrix for the chemical species based on the learned vector embeddings. The reddish color suggests similar chemical elements in the crystal environment.

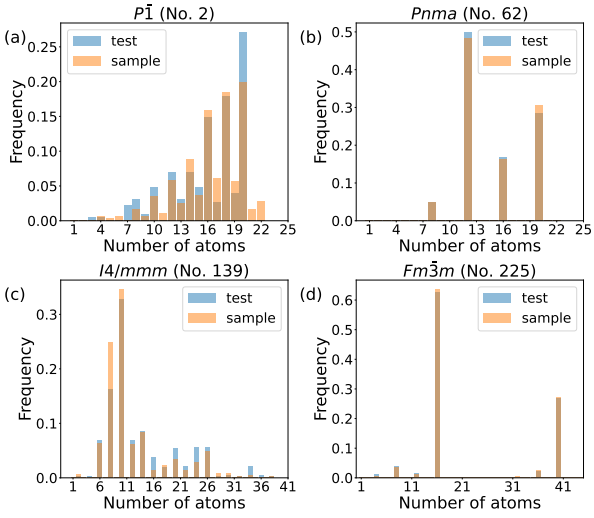


FIG. 4. The histogram for *total number* of atoms in the unit cell for several space groups in the test dataset and in the generated samples.

off positions greatly reduces the space of crystalline materials. We benchmark the `CrystalFormer` model on standard crystal generation tasks and show it exhibits great efficiency and generalization ability compared to existing models. Finally, we discuss its potential applications in novel materials discovery. We have released the codes and trained model at [32].

II. CRYSTALFORMER

To exploit the space group symmetry of the crystal, we focus on the Wyckoff positions of symmetry-inequivalent atoms. Wyckoff letters follows the alphabetical order, where "a" stands for the positions with the highest order of symmetry for the given space group. Later letters in the alphabet indicate more general positions with reduced site symmetries. Note that the information of the space group number and Wyckoff letter fully determine the multiplicities. In cases where the atom positions are fully fixed by the Wyckoff letter, we will also consider the remaining fractional coordinates, e.g. the x-coordinate of the oxygen atoms in the $CaCO_3$ example shown in Fig. 1, To generate crystals, one samples the Wyckoff letter, chemical element, and fractional coordinates of each atom sequentially. The sampling procedure starts from special higher symmetry sites with smaller multiplicities and then goes on to general lower symmetry regions with larger multiplicities.

With these considerations, we define a crystal data as $\boldsymbol{C} = \{\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}\}$. Here $\boldsymbol{W} = [w_1, w_2, \ldots, w_n]$ are Wyckoff letters and $\boldsymbol{A} = [a_1, a_2, \ldots, a_n]$ are chemical species. Here, n stands for the number of symmetrically inequivalent atoms in the conventional unit cell. For example, as shown in Fig. 1(b) one has $n = 3$ for $CaCO_3$. Explicitly including the Wyckoff letter in the generative modeling is the key of the present work. Next, $\boldsymbol{X} = [(x_i, y_i, z_i)] \in \mathbb{R}^{n\times 3}$ are the fractional coordinates of symmetrically inequivalent atoms. Lastly, $\boldsymbol{L} = [a, b, c, \alpha, \beta, \gamma]$ denotes the lattice parameters of the conventional unit cell of the material.

The central quantity to focus on is the conditional probability of a crystal $\boldsymbol{C}$ given the space group number $g \in [1, 230]$: $p(\boldsymbol{C}|g)$. Since the space group is a fundamental characterization for crystalline materials, g is a key control variable that greatly simplifies the distribution over the entire crystal materials space. In practical applications, the space group can be

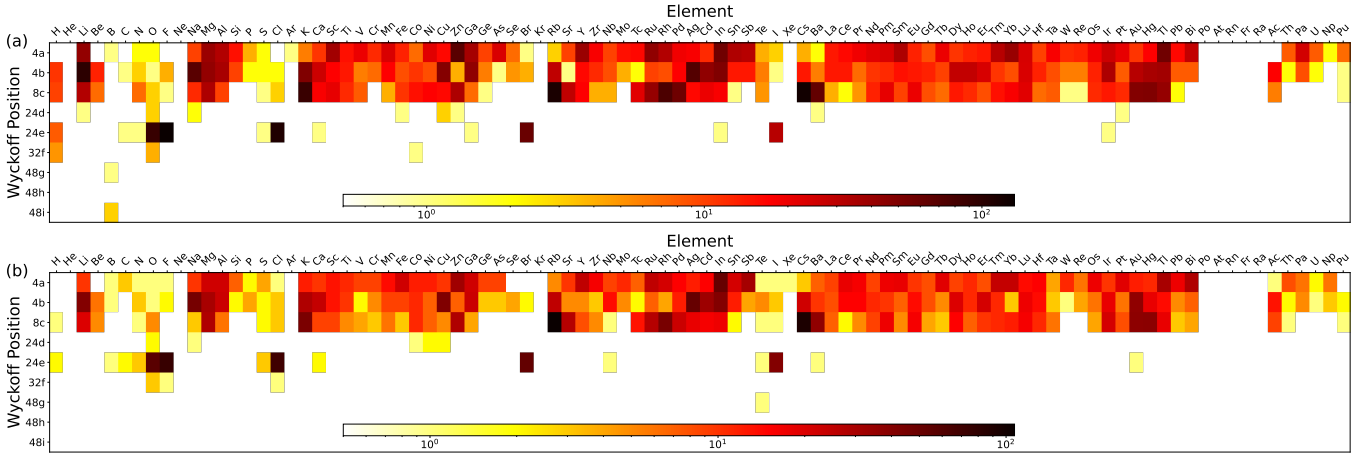


FIG. 5. The heat map for chemical elements and Wyckoff positions of (a) the test dataset and (b) generated samples for the $Fm\bar{3}m$ space group (No. 225).

considered separately or even predicted [33, 34].

We express the space group conditioned probability distribution of crystals as an autoregressive product of conditional probabilities

$$\begin{aligned} p(\boldsymbol{C}|g) \;=\; & p(w_1|g)\times \\ & p(a_1|g,w_1)\times \\ & p(x_1|g,w_1,a_1)\times \\ & p(y_1|g,w_1,a_1,x_1)\times \\ & p(z_1|g,w_1,a_1,x_1,y_1)\times\cdots\times \\ & p(\boldsymbol{L}|g,w_1,a_1,x_1,y_1,z_1\ldots,w_n,a_n,x_n,y_n,z_n). \end{aligned} \quad (1)$$

At first sight, it may appear unnatural to employ an autoregressive model for crystals since there is no obvious order for atoms in the unit cell. However, the sequential nature of Wyckoff positions suggests a natural way to arrange symmetrically inequivalent atoms in an alphabetical order of the Wyckoff letters. During sampling, the autoregressive model places the atoms one by one into the unit cell, starting from the special position with high symmetry to the general position with the lowest symmetry, see Fig. 1(c).

We model the conditional probability of the Wyckoff letters $\boldsymbol{W}$ and chemical species $\boldsymbol{A}$ as categorical distributions. A mixture of von-Mises distribution is employed to model the conditional probability of the factional coordinates $\boldsymbol{X}$. For Wyckoff positions with multiplicities greater than one, we only consider the first of fractional coordinates that appear in the international tables for crystallography [35]. Lastly, we model the conditioned distribution of lattice parameters as a Gaussian mixture model.

We build `CrystalFormer` using the vanilla decoder-only transformer [36] to model the probability distribution of crystalline materials Eq. (1). The transformer-based autoregressive model is a favorable choice to capture these prominent discrete characteristics of crystals. We feed atomistic features into an autoregressive transformer. The space group number g is the first input to the transformer. The remaining inputs alternate between the Wyckoff letter, chemical species, and fractional coordinates of each atom. One can go through the table of Fig. 1(b) in a raster order to collect these atomistic features. We feed vector embeddings of the space group number, Wyckoff letter, and the chemical species input to the `CrystalFormer`. In particular, we concatenate the vector embedding of g to all other inputs since it is the key control variable for the crystal generation. In addition, we have also provided the multiplicity of each Wyckoff position as an additional feature. The multiplicity can be easily inferred from the space group and the Wyckoff letters. We feed the fractional coordinates as Fourier features into the transformer so that the model preserves the periodicity of the unit cell [13, 37] We pad the atom sequence up to a maximum length and take the shifted output as parameters of the conditional probability distribution Eq. (1), see Fig. 1(b). At the location of the first padding atom, we predict the lattice parameters.

We implement a number of constraints in the model to further reduce the phase space of the model. First, the Wyckoff letters should be valid for the given space group. For example, for the space group number 167 the Wyckoff letters go from 'a' up to 'f'. Second, we require that the Wyckoff letters w_i follow alphabetical order in the sequence [38]. Lastly, the Wyckoff positions with no continuous degree of freedom (such as 'a', 'b', and 'd' positions in the space group 167) can only be occupied once. Those constraints are implemented by tuning the logit biases of Wyckoff letters [39, 40].

The design of `CrystalFormer` focuses mostly on the space group symmetries since we believe they are the most crucial ones for crystalline materials. This design decision significantly impacts the treatment of other symmetries. First, it is often possible to place the origin of the unit cell at the inversion center of the specified space group. The chosen origin naturally fixes the continuous translation invariance of fractional coordinates. Second, by only considering symmetry-inequivalent atoms and labeling them with Wyckoff letters, one fixes most of the permutation invariance in the representation. For those Wyckoff positions with continuous degrees of freedom, there may be multiple symmetry-

inequivalent atoms with the same Wyckoff letters. We arrange these atoms according to the lexicographic order of fractional coordinates [41] in the sequence. Note that in a crystal environment, the same type of atoms occupying different Wyckoff positions could be regarded as distinguished particles as they generally have different site symmetry. Lastly, the periodicity of the fractional coordinates is respected since they are treated as variables under the von-Mises distribution.

The `CrystalFormer` is trained by minimizing the negative log-likelihood over training dataset

$$\mathcal{L} = -\mathop{\mathbb{E}}_{\boldsymbol{C},g}\left[\ln p(\boldsymbol{C}|g)\right]. \tag{2}$$

Writing out $p(\boldsymbol{C}|g)$ according to Eq. (1), the objective function contains the negative log-likelihood of discrete variables such as Wyckoff letters $\boldsymbol{W}$ and chemical species $\boldsymbol{A}$, as well as continuous variables such as fractional coordinates $\boldsymbol{X}$ and the lattice parameters $\boldsymbol{L}$. In the objective function, for continuous variables $\boldsymbol{X}, \boldsymbol{L}$ we consider only active ones that are not fixed by the space groups and Wyckoff letters. In this way, those special fractional coordinates (e.g. 0, $\frac{1}{4}$) and lattice parameters (e.g. 90°, 120°) do not contribute to the loss function.

To sample crystals from the `CrystalFormer`, one needs to specify a space group number and a number of possible chemical elements. The `CrystalFormer` samples the atoms one by one, starting from more symmetric specific positions with lower multiplicities till less symmetric general positions with larger multiplicities. We use the information of the space group and Wyckoff letter to control the sampling of fractional coordinates. By applying the symmetry projection to the sampled fractional coordinate, one ensures the generated fractional coordinates are compatible with the Wyckoff positions. One can also mask out the logits of chemical species so that only a number of selected elements will be sampled. The number of symmetrically inequalivent atoms may fluctuate in the sampling procedure. Once one has sampled a padding atom, the model predicts the lattice parameters under the space group constraint. Moreover, we introduce a temperature parameter T in the sample distribution $p(\boldsymbol{C}|g)^{1/T}$. With $T < 1$ we will draw samples from a sharper distribution, while $T > 1$ gives more diversity in the generated samples.

III. RESULTS AND EVALUATIONS

We train the `CrystalFormer` using the MP-20 dataset [11] which is a popular dataset that represents a majority of experimentally known crystalline materials at ambient conditions with no more than 20 atoms in the primitive unit cell. The training dataset contains 27136 crystal structures. The subdivision of the training samples according to the space group has greatly reduced the number of samples in each space group category. On top of that, the distribution of training samples is quite uneven among the space groups, which reflects the imbalance distribution of crystals over space groups in Nature [30]. In fact, there is no training data in 61 out of 230 space groups as shown in Fig. 6. Nevertheless, we still employ the MP-20 as the training set so that the performance of the model can be more easily gauged with the others in the literature.

Figure 2 shows a breakup of the learning curves for the Wyckoff position, chemical species, fractional coordinates, and lattice parameters. Next, we will select the model checkpoint with the lowest total validation loss and examine its learned features. Then, we will use the model to generate 1000 crystal samples for each of the 230 space groups and evaluate their validity, stability, and novelty. These results highlight the benefits of built-in space group symmetries in the model and point to possible future applications of the `CrystalFormer` model.

A. Learned features

Figure 3 visualizes the cosine similarity of the learned vector embedding of the chemical species. Red colors in the figure indicate similar chemical species identified by the model. One clearly sees the periodicity shows up as off-diagonal stripes. Moreover, there are visible clusters for Lanthanide elements (La-Lu). The plot also suggests the similarity between the lanthanides and other rare-earth elements (Y and Sc). Being able to learn chemical similarities from data [19, 42–46] is an encouraging signal that the model is picking up atomic physics to be able to generate reasonable crystal structures. In particular, the features shown in Fig. 3 is strikingly similar to the similarity map constructed purposely based on substitution pattern [42] which was later used for substitution-based material discovery [47].

Figure 4 presents the histogram of the total number of atoms in the conventional unit cell for several space groups. One sees a nice agreement between the atom number distribution in the test dataset and the generated samples. In addition, it appears that space group g is the key latent variable that decomposes the multi-modal atom number distribution of crystals. This is understandable because the number of atoms is determined by the sum of the multiplicities of occupied Wyckoff positions. Therefore, the space group symmetry is a key control variable for the atom number distribution. Building Wyckoff positions information into the `CrystalFormer` model architecture removes the necessities of querying the training data to find out the number of atoms for a targeted space group [16].

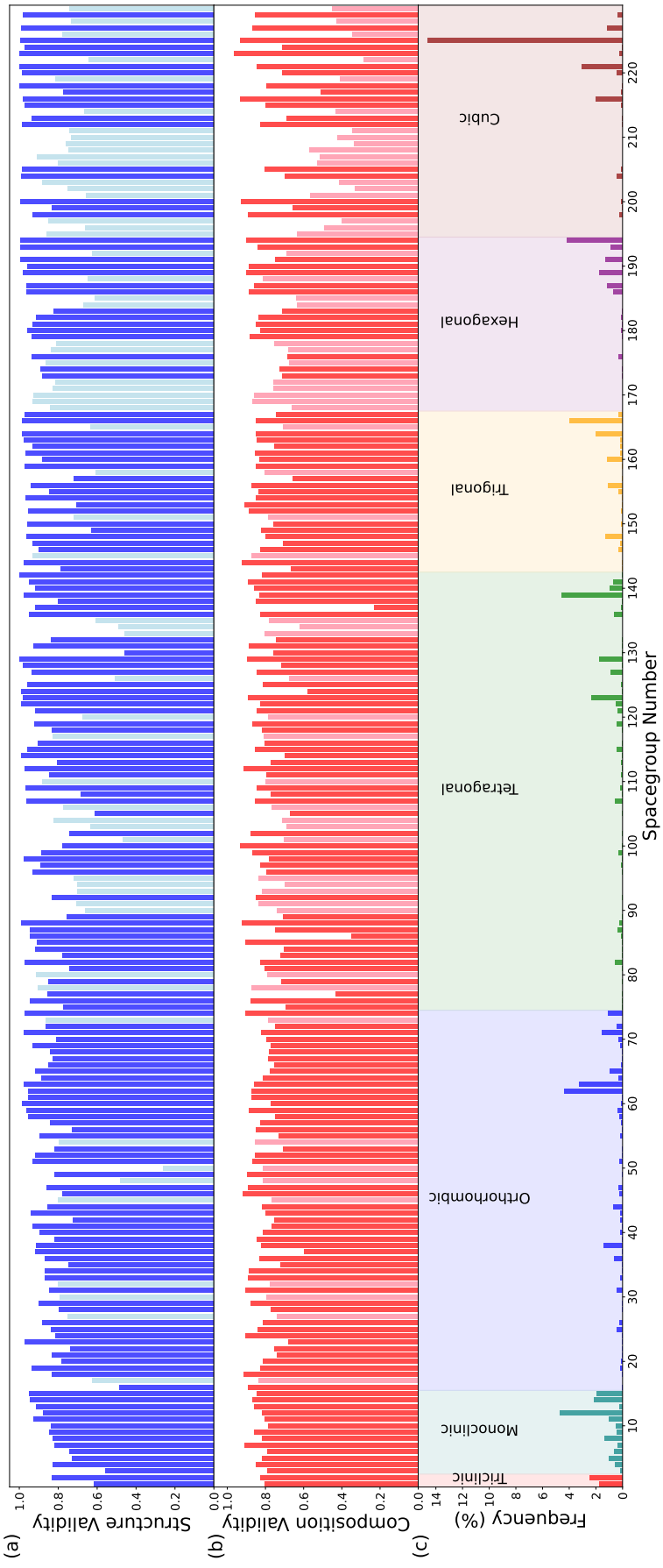


FIG. 6. (a) Structure validity histogram, (b) Composition validity histogram (c) frequency of space groups in the training dataset. The histograms in light colors in (a) and (b) are for those space group where there is no training data. A dataset of 1000 generated samples per space group was curated for evaluation.

Figure 5 shows heat maps of Wyckoff positions and chemical species for the $Fm\bar{3}m$ space group (No. 225). First of all, one sees that most atoms occupy those special Wyckoff positions with high symmetries. Moreover, the heat map reveals interesting solid-state chemistry knowledge as it suggests where each atom tends to appear in a crystal. For example, there are vertical stripes at the locations of inert elements (He, Ne, Ar...) as they are rare in crystalline materials. Moreover, one sees that oxygen and halogen elements (F, Cl, Br, I) appear quite often in the Wyckoff position "24e", which is a consequence of their electronegativities. Overall, we see the `CrystalFormer` has learned these key motifs for generating crystalline materials. On the other hand, one sees that several Wyckoff locations of the hydrogen are missing in the generated samples compared to the test dataset. We believe that is due to that the hydrogen element takes only about 0.4% in the training data for the $Fm\bar{3}m$ space group. Collecting more data with better coverage of elements will be crucial to further boost the performance of the current model.

There has been a long history of summarizing empirical or statistical chemistry rules encoded in materials data and then use them to instruct search of crystal structures [42, 48–52]. Our analysis shows that `CrystalFormer` also ingests a number of chemical intuitions in the training data for generative modelling. Since training procedure compresses the chemistry knowledge into a neural network model with tractable likelihoods, it is possible to employ `CrystalFormer` in the Monte Carlo or evolution strategy search of crystal structures, besides as a probabilistic generative model.

B. Validity and novelty of generated samples

Figure 6 illustrates the structure and compositional validity of generated samples across all 230 space groups. Following the Ref. [53], a structure meets the validity criteria if the shortest atomic distance exceeds 0.5 Å, a lenient standard. Composition validity requires charge neutrality as computed by SMACT [54]. This is, however, an overly stringent criterion since the composition validity of the training set is only around 90% by this measure [11]. Note that the `CrystalFormer` is able to generate reasonable samples even for those space groups without any training data.

Table II reports the validity of generated samples for selected space groups in each of the seven crystal systems. To ensure the numbers are representative, we chose the space group to be the one with the most training data for each crystal system. One sees that the model performs better for more symmetric space groups. This is a nice feature that complements existing crystal generative models, which mostly have difficulties in generating highly symmetric structures. As a reference, we also list the validity of the generated samples suppose one treats the crystals as if they are all in the $P1$ space group (No. 1) with only translational symmetry. One sees the structure validity scores of $P1$ space group improves compared to the one shown in Figure 6 due to increased training samples.

The second part of Table II shows reference results in the literature for the same validity test. In principle, the performance of the present model should fall back to the language model approaches [18, 20]. The remaining gap may be due to details such as the including the header line in the crystallographic information files (CIF), specific sampling strategy of language models, or the additional post-selection of samples [20]. In the table, the DiffCSP++ [26] is the only alternative model that exploits the space group symmetry in the generation process. However, different from us, the DiffCSP++ model does not predict Wycokff position as a part of generation process. Therefore, DiffCSP++ needs to search for template structure in the training set for generation, which may limit its generality. Besides works listed in Table II that reported validity scores in a comparable settings, Ref. [19] has conditioned the generation of CIF on the space group symbols in a language model setting. Ref. [16] considered space group conditioned crystal generation using a fine-tuned generative model with space group labels. Neither approach provides exact constraints on the space group, which could yield problematic structures for large systems and highly symmetric space groups. Ref. [23] considered generating symmetric crystals in their Wyckoff representation. However, the model does not consider fractional coordinates and lattice parameters, so it requires a subsequent computational search to completely determine the crystal structure. Sec. IV B contains a more in-depth discussion of the relationship of the present approach with related works.

Figure 7(a) shows the validity of generated samples as a

TABLE II. Validity of generated crystal structure for representative space groups. Training samples count the number of samples in the training set.

Space group	**Crystal system**	**Training samples**	**Validity** (%) ↑	
			Struc.	Comp.
2	Triclinic	676	83.10	83.0
12	Monoclinic	1273	87.70	81.80
62	Orthorhombic	1187	95.50	87.20
139	Tetragonal	1233	97.70	83.40
166	Trigonal	1076	98.50	85.0
194	Hexagonal	1129	99.40	89.90
225	Cubic	3960	99.60	93.50
1	Triclinic	27136	91.40	80.20
Autoregressive models				
PGSchNet [55]			99.65	75.96
LM-CH (character-level tokenization) [18]			84.81	83.55
LM-AC (atom coordinate-level tokenization) [18]			95.81	88.87
Crystal-LLM [20]			96.5	86.3
Diffusion models				
CDVAE [11]			100.0	86.70
DiffCSP [13]			100.0	83.25
DiffCSP++ [26]			99.94	85.12
UniMat-Large [15]			97.2	89.4

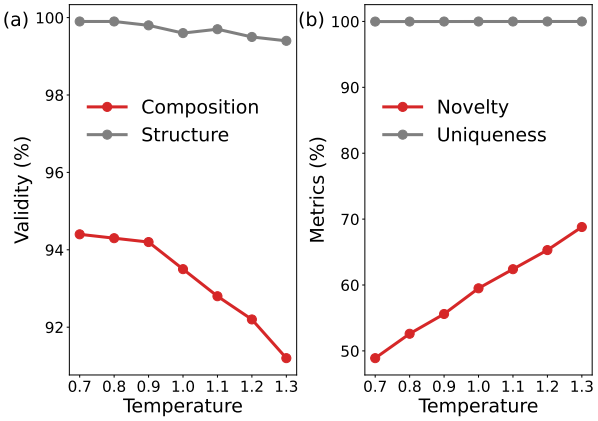


FIG. 7. (a) Structure and composition validity, (b) novelty and uniqueness of generated samples evaluated according to [11] for the $Fm\bar{3}m$ space group.

function of sampling temperature. One clearly sees that reduced temperature $T < 1$ increases the validity of samples at the cost of reducing the diversity [20]. Figure 7(b) shows the novelty and uniqueness evaluated on 1000 generated samples with temperature. Novelty quantifies the proportion of new structures in the generated samples that were unseen in the training dataset. Uniqueness represents the percentage of distinct, non-redundant structures among the generated samples [56]. One sees that across different temperatures the uniqueness remains high, indicating the model does not collapse to a mode that produces duplicated samples. On the other hand, the model produces close to 70% novel material as temperature increases, which nicely demonstrates modal covering behavior of the maximum likelihood estimation training [57]. Having a model distribution broader than the span of the dataset is crucial for material discovery.

C. Stability of generative samples

Figure 8(a) shows the histogram of the log-likelihood of generated and test samples show nice agreement. We also visualize structures of a few generated samples which deems to be very likely, typical, and unlikely according to their log-likelihoods. Interestingly, we did not observe a clear correlation between the model likelihood and energy of the structure.

Fig. 8(b) shows the energy above hull based on energies calculated with density functional theory (DFT). One sees the distributions of unrelaxed and relaxed structures are very similar, suggesting that the generated structures are already very close to local minima of DFT calculations. For the given computational budget detailed in the Appendix B, we manage to reach convergence for 973 out of 1000 structures in the DFT relaxation. All of the structures retains the same space group symmetry after the DFT relaxation [59]. With its default parameters, `StructureMatcher` from pymatgen [60] is able to match all 973 structures before and after relaxation. The average root mean squared displacement (RMSD) normalized by $\sqrt[3]{V/N}$ computed for these matched structures is 0.0061, which confirms the generated samples are indeed very close to DFT local minima. Figure 8(b) shows a pronounced peak around $E_{\text{hull}} \approx 0.1$eV/atom showing the model indeed generates a large fraction of candidates of metastable materials [58]. Finally, we found 21 relaxed samples are stable by the criteria $E_{\text{hull}} < 0$ and summarize them in Table S2 of appendix B. Most of the found stable materials only have atoms occupying the first few Wyckoff positions. Among these 21 structures, we found 10 are new crystalline materials that are not contained in the Materials Project database [61].

IV. DISCUSSIONS

A. Applications

With the abilities demonstrated in Sec. III, we believe that `CrystalFormer` is ready to fit into the existing materials discovery workflow in the following two ways.

1. *Initialization for structure search*

Crystal structure search relies heavily on utilizing space group symmetries. For example, it is a common practice for crystal structure prediction software [62–66] and crystal structure search [67, 68] to place atoms at Wyckoff positions to prepare symmetric initial states for subsequent optimization.

However, such initialization approach faces combinatorial difficulty as the number of chemical species and atoms in the unit cell grow. The `CrystalFormer` is ready to act as a drop-in replacement of random structure initialization. In this way, one bypasses the curse-of-dimensionality of exact enumeration with a data-driven probabilistic approach. Moreover, the ability of `CrystalFormer` to generate diverse and close to metastable structures can greatly reduce computational costs of downstream ab initio calculations.

2. *Mutation of existing structures*

Mutatation of known crystals is a prominent approach to materials discovery. For example, one can employ machine machine-learned force field to relax crystal structures [5, 69–71] after element substitutions. In the lens of generative modeling, the machine learning force field can be regarded as the energy-based model or Boltzmann machines. A potential drawback of exploring materials space with an energy-based model is the slow mixing or even ergodicity issue posed by the rough landscape of the potential energy surface. In this sense, element substitutions provide a variety of initial seeds, compensating the limitation potential energy surface-based exploration.

Having an alternative measure of crystal likelihood other than the potential energy surface opens a way to employ the model likelihood as a guide for structure search. For example, with a trained `CrystalFormer`, one can perform Markov

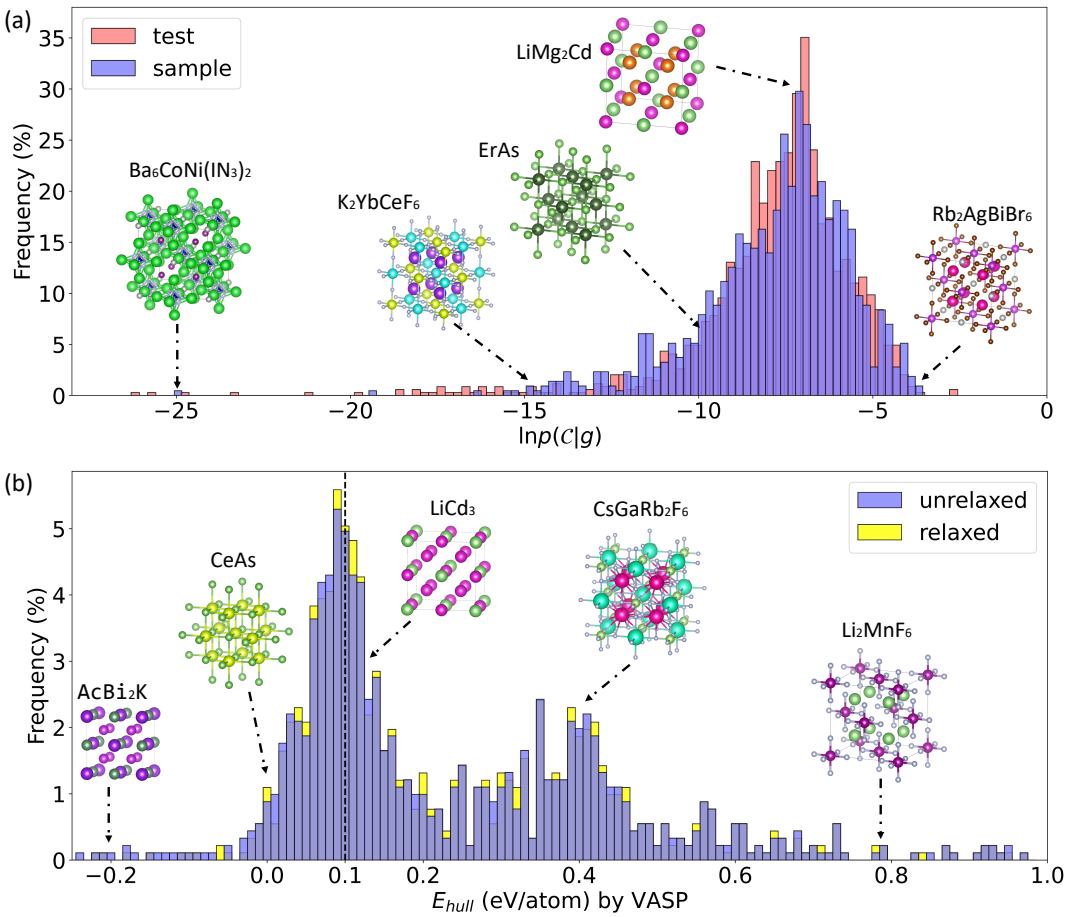


FIG. 8. (a) The log-likelihood histogram for 1000 samples in the $Fm\bar{3}m$ space group and the test dataset. (b) The histograms of energy above the convex hull. The dashed line denotes the metastable criterion because materials with $E_{\text{hull}} < 0.1$ eV/atom are usually metastable and have potential utilities in practice [58]. The insets visualize the crystal structure of a few generated samples. $Rb_2AgBiBr_6$, $LiMg_2Cd$, and $LiCd_3$ are in train dataset. ErAs is in the validation dataset. CeAs is in the test dataset. K_2YbCeF_6, $Ba_6CoNi(IN_3)_2$, $CsGaRb_2F_6$, Li_2MnF_6, $AcBi_2K$ are not in the MP-20 or Materials Project database.

chain Monte Carlo (MCMC) random walk in the crystalline materials starting from an existing crystal structure. At each step of the random walk, one proposes an element substitution, atom shift, or lattice deformation and accept or reject the proposal according to the model likelihood. In the long time limit, one will sample from the model distribution. However, during the burn-in phase of such MCMC sampling, the generated samples will be similar to the starting material, which may be a desired feature in certain cases.

B. Related works

Crystal generative models have been explored using variational autoencoder [23, 72], generative adversarial networks [12, 73], normalizing flows [74–76], diffusion models [11–16, 21, 22, 26], GFlowNet [24, 25], and autoregressive models [17–19, 55, 77, 78]. In these autoregressive models, one either use atomistic features [55, 77, 78] or use pure text tokens [17–20]. However, with the introduction of specialized tokens for crystals, the boundary between the two is blurred.

The `CrystalFormer` is most closely related to the autoregressvie generative model originally designed for molecules [55, 77, 78]. However, instead of predicting the relative distances of atoms, we predict the Wyckoff positions of atoms in the unit cell. Having the luxury of the space group symmetry for crystals provides strong hints on where to put the atoms in the unit cell and greatly simplifies the design around rotational equivariance. On the other hand, compared to Refs. [17–20] which treat text descriptions of crystals using autoregressive language models, we are only dealing with more concise and essential atomistic representation of crystals, which leads to smaller model size and faster sampling speed. Fast generation speed is not only a welcoming feature but also will be crucial for intelligent exploration of materials space based on combinations of probabilistic generation and post-selection, backtracking, and searching techniques [79]. More importantly, `CrystalFormer` guarantees space group constraints since these symmetries are baked in the model architecture rather than learned as statistical correlation from

text data. In a sense, the present work employs intrinsic mathematical (as opposed to natural) language to incorporate the symmetry principle in the generative modeling of crystals.

As a side remark, the Wyckoff position features have been used in machine learning models for materials property prediction [80, 81]. Incorporating space group information in the encoder-only transformer models may also boost their property prediction performance [82, 83] as suggested by the related study [84].

V. OUTLOOK

Crystal structure prediction has long been the dream of computational material sciences [85]. `CrystalFormer` integrates exact symmetry principles from math and empirical chemical intuitions from data into one framework. Probabilistic generative modeling of crystalline materials using `CrystalFormer` opens the way to many future innovations in materials discovery.

Precisely controlling the space group in the generative model of crystalline materials is not only a highly desired feature but also greatly simplifies the task. An obvious future direction is to scale up the model as well as the training dataset, especially curating a dataset with better coverage of space groups. Note that the MP-20 dataset has by no means exhausted all available crystalline material [16, 19]. The transformer-based generative model is ready to be scaled up with more training data, in the same fashion as large language models [86]. Given similar model architectures, the idea of generative pretraining of a foundational model for material generation is appealing. When scaling up the model it will be interesting to note the possible appearance of neural scaling law [87] as it has also been showing up in other contexts of atomistic modeling [88].

The model architecture is still open to further improvement to better serve the purpose of material discovery, First of all, to better facilitate data efficiency learning and structure phase transitions-related applications, it will be useful to further exploit the group-subgroup structure [89] in the model architecture. Second, besides chemical elements, it will be useful to constrain the stoichiometry of generated materials. This may be achieved by tuning the logit bias or combining the present probabilistic generation approaches with traditional enumeration approaches [90]. Lastly, it is worth exploring to use `CrystalFormer` as the base distribution in the flow model and employ symmetry-persevering transportation's to further adjust the atoms coordinates and unit cells [10, 26], which mimics a symmetry-constrained relaxation process [91].

Conditioned crystal structure generation based on material properties [16, 78] and experimental measurements [92] are both straightforward with `CrystalFormer`. One can either extend the space group embedding or employ the full encoder-decoder transformer architecture [36] for conditioned crystal generation. The first step of property-guided generation could take into account of synthesizability and stability of materials. In the future, proposing materials with multiple properties is a promising direction for material generative models. To achieve that goal, curating a high-quality material dataset with either computed or measured properties is crucial. Needless to say, to close the loop of material discovery, one will need to carry out experimental verification for the generated materials.

ACKNOWLEDGMENTS

We thank Han Wang, Lin Yao, Linfeng Zhang, Chen Fang, Yanchao Wang, Qi Yang, Shigang Ou, Xinyang Dong, and Quansheng Wu for useful discussions. This project is supported by the National Natural Science Foundation of China under Grants No. T2225018, No. 92270107, No. 12188101, No. T2121001, and No. 12034009 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grants No. XDB0500000 and No. XDB30000000.

[1] S. M. Woodley and R. Catlow, Crystal structure prediction from first principles, Nature Materials **7**, 937 (2008).
[2] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, Structure prediction drives materials discovery, Nature Reviews Materials **4**, 331 (2019).
[3] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Central Science **4**, 268 (2018), pMID: 29532027.
[4] B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science **361**, 360 (2018).
[5] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, Nature **624**, 80 (2023).
[6] C. Chen, D. T. Nguyen, S. J. Lee, N. A. Baker, A. S. Karakoti, L. Lauw, C. Owen, K. T. Mueller, B. A. Bilodeau, V. Murugesan, and M. Troyer, Accelerating computational materials discovery with artificial intelligence and cloud high-performance computing: from large-scale screening to experimental validation, (2024), arXiv:2401.04070 [cond-mat.mtrl-sci].
[7] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, Gpt-4 technical report, (2023), arXiv:2303.08774 [cs.CL].
[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, Zero-shot text-to-image generation, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 8821–8831.
[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF conference on computer vision*

and pattern recognition (2022) pp. 10684–10695.
[10] J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, and G. Grigoryan, Illuminating protein space with a programmable generative model, Nature **623**, 1070 (2023).
[11] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, Crystal diffusion variational autoencoder for periodic material generation, (2021), arXiv:2110.06197 [cs.LG].
[12] Y. Luo, C. Liu, and S. Ji, Towards symmetry-aware generation of periodic materials, (2023), arXiv:2307.02707 [cs.LG].
[13] R. Jiao, W. Huang, P. Lin, J. Han, P. Chen, Y. Lu, and Y. Liu, Crystal structure prediction by joint equivariant diffusion, (2023), arXiv:2309.04475 [cond-mat.mtrl-sci].
[14] S. Zheng, J. He, C. Liu, Y. Shi, Z. Lu, W. Feng, F. Ju, J. Wang, J. Zhu, Y. Min, H. Zhang, S. Tang, H. Hao, P. Jin, C. Chen, F. Noé, H. Liu, and T.-Y. Liu, Towards predicting equilibrium distributions for molecular systems with deep learning, (2023), arXiv:2306.0544 [physics.chem-ph].
[15] M. Yang, K. Cho, A. Merchant, P. Abbeel, D. Schuurmans, I. Mordatch, and E. D. Cubuk, Scalable diffusion for materials discovery, (2023), arXiv:2311.09235 [cs.LG].
[16] C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, *et al.*, Mattergen: a generative model for inorganic materials design, (2023), arXiv:2312.03687 [cond-mat.mtrl-sci].
[17] H. Xiao, R. Li, X. Shi, Y. Chen, L. Zhu, X. Chen, and L. Wang, An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning, Nature Communications **14**, 7027 (2023).
[18] D. Flam-Shepherd and A. Aspuru-Guzik, Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files, (2023), arXiv:2305.05708 [cs.LG].
[19] L. M. Antunes, K. T. Butler, and R. Grau-Crespo, Crystal structure generation with autoregressive large language modeling, (2023), arXiv:2307.04340 [cond-mat.mtrl-sci].
[20] N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, and Z. Ulissi, Fine-tuned language models generate stable inorganic materials as text, (2024), arXiv:2402.04379 [cs.LG].
[21] X. Luo, Z. Wang, P. Gao, J. Lv, Y. Wang, C. Chen, and Y. Ma, Deep learning generative model for crystal structure prediction, (2024), arXiv:2403.10846 [cond-mat.mtrl-sci].
[22] C.-Y. Ye, H.-M. Weng, and Q.-S. Wu, Con-cdvae: A method for the conditional generation of crystal structures, (2024), arXiv:2403.12478 [cond-mat.mtrl-sci].
[23] R. Zhu, W. Nong, S. Yamazaki, and K. Hippalgaonkar, Wycryst: Wyckoff inorganic crystal generator framework, (2023), arXiv:2311.17916 [cond-mat.mtrl-sci].
[24] M. AI4Science, A. Hernandez-Garcia, A. Duval, A. Volokhova, Y. Bengio, D. Sharma, P. L. Carrier, M. Koziarski, and V. Schmidt, Crystal-gfn: sampling crystals with desirable properties and constraints, (2023), arXiv:2310.04925 [cs.LG].
[25] T. M. Nguyen, S. A. Tawfik, T. Tran, S. Gupta, S. Rana, and S. Venkatesh, Hierarchical GFlownet for crystal structure generation (2024).
[26] R. Jiao, W. Huang, Y. Liu, D. Zhao, and Y. Liu, Space group constrained crystal generation, (2024), arXiv:2402.03992 [cs.LG].
[27] A. R. Oganov and C. W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, The Journal of chemical physics **124** (2006).
[28] D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, and A. Walsh, Computational screening of all stoichiometric inorganic materials, Chem **1**, 617 (2016).
[29] M. Glazer, G. Burns, and A. Glazer, *Space Groups for Solid State Scientists* (Elsevier Science, 2012).
[30] V. S. Urusov and T. N. Nadezhina, Frequency distribution and selection of space groups in inorganic crystal chemistry, Journal of Structural Chemistry **50**, 22 (2009).
[31] R. E. Marsh, P1 or P1¯? Or something else?, Acta Crystallographica Section B **55**, 931 (1999).
[32] See https://github.com/deepmodeling/CrystalFormer for code and model checkpoint.
[33] Y. Zhao, Y. Cui, Z. Xiong, J. Jin, Z. Liu, R. Dong, and J. Hu, Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions, ACS Omega **5**, 3596 (2020), pMID: 32118175.
[34] D.-Y. Wang, H.-F. Lv, and X.-J. Wu, Crystallographic groups prediction from chemical composition via deep learning, Chinese Journal of Chemical Physics **36**, 66 (2023).
[35] T. Hahn, U. Shmueli, and J. W. Arthur, *International tables for crystallography*, Vol. 1 (Reidel Dordrecht, 1983).
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017).
[37] P. Wirnsberger, A. J. Ballard, G. Papamakarios, S. Abercrombie, S. Racanière, A. Pritzel, D. J. Rezende, and C. Blundell, Targeted free energy estimation via learned mappings, Journal of Chemical Physics **153**, 144112 (2020).
[38] We follow the convention that capital letters appear *after* lower case letters. This handles the edge case of the No. 47 space group *Pmmm* whose Wyckoff positions reach 'A'. This rule also demands that the padded atoms appear at the end of the sequence with Wyckoff letter 'X'.
[39] OpenAI, Using logit bias to alter token probability with the openai api, OpenAI Help Center, retrieved March 14, 2024.
[40] H. Xie, L. Zhang, and L. Wang, m^* of two-dimensional electron gas: A neural canonical transformation study, SciPost Phys. **14**, 154 (2023).
[41] E. Parthé and L. M. Gelato, The standardization of inorganic crystal-structure data, Acta Crystallographica Section A **40**, 169 (1984).
[42] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining, New Journal of Physics **18**, 093011 (2016).
[43] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, Learning atoms for materials discovery, Proceedings of the National Academy of Sciences **115**, E6411 (2018).
[44] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials **31**, 3564 (2019).
[45] D. Zhang, H. Bi, F.-Z. Dai, W. Jiang, L. Zhang, and H. Wang, Dpa-1: Pretraining of attention-based deep potential model for molecular simulation, (2022), arXiv:2208.08236 [physics.chem-ph].
[46] A. Y.-T. Wang, M. S. Mahmoud, M. Czasny, and A. Gurlo, Crabnet for explainable deep learning in materials science: bridging the gap between academia and industry, Integrating Materials and Manufacturing Innovation **11**, 41 (2022).
[47] H.-C. Wang, S. Botti, and M. A. L. Marques, Predicting stable

crystalline compounds using chemical similarity, npj Computational Materials **7**, 12 (2021).
[48] L. Pauling, The principles determining the structure of complex ionic crystals, Journal of the American Chemical Society **51**, 1010 (1929).
[49] V. Goldschmidt, Crystal structure and chemical constitution, Transactions of the Faraday Society **25**, 253 (1929).
[50] D. Pettifor, Structure maps for. pseudobinary and ternary phases, Materials Science and Technology **4**, 675 (1988).
[51] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics, Nature Materials **5**, 641 (2006).
[52] Z. Allahyari and A. R. Oganov, Coevolutionary search for optimal materials in the space of all possible compounds, npj Computational Materials **6**, 55 (2020).
[53] C. J. Court, B. Yildirim, A. Jain, and J. M. Cole, 3-d inorganic crystal structure generation and property prediction via representation learning, Journal of Chemical Information and Modeling **60**, 4518 (2020), pMID: 32866381.
[54] D. W. Davies, K. T. Butler, A. J. Jackson, J. M. Skelton, K. Morita, and A. Walsh, Smact: Semiconducting materials by analogy and chemical theory, Journal of Open Source Software **4**, 1361 (2019).
[55] N. Gebauer, M. Gastegger, and K. Schütt, Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules, in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).
[56] https://github.com/sparks-baird/matbench-genmetrics.
[57] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016) http://www.deeplearningbook.org.
[58] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and G. Ceder, The thermodynamic scale of inorganic crystalline metastability, Science Advances **2**, e1600225 (2016).
[59] Due to merging of Wyckoff positions that are occupied with the same elements, 26 structures actually belong to the $Pm\bar{3}m$ space group (No. 221).
[60] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science **68**, 314 (2013).
[61] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials **1**, 011002 (2013).
[62] C. J. Pickard and R. J. Needs, Ab initio random structure searching, Journal of Physics: Condensed Matter **23**, 053201 (2011).
[63] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Calypso: A method for crystal structure prediction, Computer Physics Communications **183**, 2063 (2012).
[64] A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, New developments in evolutionary structure prediction algorithm uspex, Computer Physics Communications **184**, 1172 (2013).
[65] P. Avery and E. Zurek, Randspg: An open-source program for generating atomistic crystal structures with specific spacegroups, Computer Physics Communications **213**, 208 (2017).
[66] S. Fredericks, K. Parrish, D. Sayre, and Q. Zhu, Pyxtal: A python library for crystal structure generation and symmetry analysis, Computer Physics Communications **261**, 107810 (2021).
[67] G. Cheng, X.-G. Gong, and W.-J. Yin, Crystal structure prediction by combining graph network and optimization algorithm, Nature Communications **13**, 1492 (2022).
[68] H.-C. Wang, J. Schmidt, M. A. L. Marques, L. Wirtz, and A. H. Romero, Symmetry-based computational search for novel binary and ternary 2d materials, 2D Materials **10**, 035007 (2023).
[69] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science **2**, 718 (2022).
[70] D. Zhang, X. Liu, X. Zhang, C. Zhang, C. Cai, H. Bi, Y. Du, X. Qin, J. Huang, B. Li, Y. Shan, J. Zeng, Y. Zhang, S. Liu, Y. Li, J. Chang, X. Wang, S. Zhou, J. Liu, X. Luo, Z. Wang, W. Jiang, J. Wu, Y. Yang, J. Yang, M. Yang, F.-Q. Gong, L. Zhang, M. Shi, F.-Z. Dai, D. M. York, S. Liu, T. Zhu, Z. Zhong, J. Lv, J. Cheng, W. Jia, M. Chen, G. Ke, W. E, L. Zhang, and H. Wang, Dpa-2: Towards a universal large atomic model for molecular and material simulation, (2023), arXiv:2312.15492 [physics.chem-ph].
[71] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, A foundation model for atomistic materials chemistry, (2024), arXiv:2401.00096 [physics.chem-ph].
[72] Z. Ren, S. I. P. Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun, *et al.*, An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties, Matter **5**, 314 (2022).
[73] Y. Zhao, M. Al-Fahdi, M. Hu, E. M. Siriwardane, Y. Song, A. Nasiri, and J. Hu, High-throughput discovery of novel cubic crystal materials using deep generative neural networks, Advanced Science **8**, 2100566 (2021).
[74] R. Ahmad and W. Cai, Free energy calculation of crystalline solids using normalizing flows, Modelling and Simulation in Materials Science and Engineering **30**, 065007 (2022).
[75] P. Wirnsberger, G. Papamakarios, B. Ibarz, S. Racanière, A. J. Ballard, A. Pritzel, and C. Blundell, Normalizing flows for atomic solids, Machine Learning: Science and Technology **3**, 025009 (2022).
[76] J. Köhler, M. Invernizzi, P. De Haan, and F. Noé, Rigid body flows for sampling molecular crystal structures, (2023), arXiv:2301.11355 [cs.LG].
[77] N. W. Gebauer, M. Gastegger, and K. T. Schütt, Generating equilibrium molecules with deep neural networks, (2018), arXiv:1810.11347 [stat.ML].
[78] N. W. A. Gebauer, M. Gastegger, S. S. P. Hessmann, K.-R. Müller, and K. T. Schütt, Inverse design of 3d molecular structures with conditional generative neural networks, Nature Communications **13**, 973 (2022).
[79] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, (2023), arXiv:2305.10601 [cs.CL].

[80] A. Jain and T. Bligaard, Atomic-position independent descriptor for machine learning of material properties, Phys. Rev. B **98**, 214112 (2018).

[81] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, and A. A. Lee, Rapid discovery of stable materials by coordinate-free coarse graining, Science Advances **8**, eabn4117 (2022).

[82] K. Yan, Y. Liu, Y. Lin, and S. Ji, Periodic graph transformers for crystal material property prediction, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 15066–15080.

[83] T. Taniai, R. Igarashi, Y. Suzuki, N. Chiba, K. Saito, Y. Ushiku, and K. Ono, Crystalformer: Infinitely connected attention for periodic structure encoding, (2024), arXiv:2403.11686 [cs.LG].

[84] A. N. Rubungo, C. Arnold, B. P. Rand, and A. B. Dieng, Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions, (2023), arXiv:2310.14029 [cs.CL].

[85] J. Maddox, Crystals from first principles, Nature **335**, 201 (1988).

[86] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, Language models are few-shot learners, Advances in neural information processing systems **33**, 1877 (2020).

[87] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, Scaling laws for neural language models, (2020), arXiv:2001.08361 [cs.LG].

[88] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gómez-Bombarelli, C. W. Coley, and V. Gadepally, Neural scaling of deep chemical models, Nature Machine Intelligence **5**, 1297 (2023).

[89] H. T. Stokes and D. M. Hatch, *Isotropy Subgroups of the 230 Crystallographic Space Groups* (WORLD SCIENTIFIC, 1989).

[90] U. Müller, *Symmetry Relationships between Crystal Structures: Applications of Crystallographic Group Theory in Crystal Chemistry*, International Union of Crystallography Texts on Crystallography (OUP Oxford, 2013).

[91] S. Cox and A. D. White, Symmetric molecular dynamics, Journal of Chemical Theory and Computation **18**, 4077 (2022).

[92] Q. Lai, L. Yao, Z. Gao, S. Liu, H. Wang, S. Lu, D. He, L. Wang, C. Wang, and G. Ke, End-to-end crystal structure prediction from powder x-ray diffraction, (2024), arXiv:2401.03862 [physics.chem-ph].

[93] H.-C. Wang, S. Botti, and M. A. L. Marques, Predicting stable crystalline compounds using chemical similarity, npj Computational Materials **7**, 12 (2021).

[94] J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, Large-scale machine-learning-assisted exploration of the whole materials space (2022), arXiv:2210.00579 [cond-mat.mtrl-sci].

[95] J. Schmidt, H.-C. Wang, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, A dataset of 175k stable and metastable materials calculated with the pbesol and scan functionals, Scientific Data **9**, 64 (2022).

[96] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[97] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B **50**, 17953 (1994).

[98] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

Appendix A: More details of `CrystalFormer`

To recap, the space group information plays a key role in the architecture, training, and sampling of `CrystalFormer`. First of all, the vector embedding of space group number g controls all subsequent outputs in the transformer that determines the Wyckoff letters, chemical species, fractional coordinates, and lattice parameters. Second, the information of the space group and Wyckoff letter are used to select active components in the fractional coordinates and lattice parameters in the loss function during training. Lastly, the space group determines the concrete meaning of in terms of multiplicities and fractional coordinates. This information is used to place right amount of atoms to appropriate locations during sampling.

1. Model architectures

Algorithm 1 summarized the model architecture of `CrystalFormer`. Training the model for 3,800 epochs with the hyperparameters shown in Table S1 takes about 13 hours on a single A100 GPU.

Algorithm 1 The `CrystalFormer` architecture

Input: Space group number g, Wyckoff letters $\boldsymbol{W} = [w_i]$, chemical elements $\boldsymbol{A} = [a_i]$, fractional coordinates $\boldsymbol{X} = [(x_i, y_i, z_i))]$ of each atom in the unit cell.

Output: Parameters for the conditional probability of Wyckoff letters $\boldsymbol{\omega}_i$, chemical element $\boldsymbol{\alpha}_i$, and fractional coordinates $\boldsymbol{\chi}_i, \boldsymbol{\upsilon}_i, \boldsymbol{\zeta}_i$ of atoms and the lattice parameters $\boldsymbol{\ell}$.

$\boldsymbol{\omega}_1 = \texttt{Net}(g)$ ▷ the logit of the first Wyckoff position is implemented as a standalone neural network.
if $\boldsymbol{W} = \emptyset$ **then**.
  return $\boldsymbol{\omega}_1$
end if
# multiplicity
for $i = 1{:}\texttt{len}(\boldsymbol{W})$ **do**
  $m_i = \texttt{mult\_table}[g, w_i]$
end for
# prepare input features into the transformer
$\boldsymbol{h}_W = [\texttt{Embed}(g), \texttt{Embed}(w_i), m_i]$.
$\boldsymbol{h}_A = [\texttt{Embed}(g), \texttt{Embed}(a_i)]$.
$\boldsymbol{h}_X = [\texttt{Embed}(g), \cos(2\pi x_i), \sin(2\pi x_i), \ldots, \cos(2\pi x_i N_f), \sin(2\pi x_i N_f)]$
$\boldsymbol{h}_Y = \ldots$
$\boldsymbol{h}_Z = \ldots$
concatenate along particle dimension
$\boldsymbol{h} = \texttt{Concatenate}(\boldsymbol{h}_W, \boldsymbol{h}_A, \boldsymbol{h}_X, \boldsymbol{h}_Y, \boldsymbol{h}_Z)$
Project $\boldsymbol{h}$ feature size to d_{model} and add position embedding
$\boldsymbol{h} = \texttt{MaskedTransformer}(\boldsymbol{h})$
Project $\boldsymbol{h}$ feature size to desired dimensions
split along particle dimension
$\boldsymbol{\omega}_i, \boldsymbol{\alpha}_i, \boldsymbol{\chi}_i, \boldsymbol{\upsilon}_i, \boldsymbol{\zeta}_i, \boldsymbol{\ell} = \texttt{Split}(\boldsymbol{h})$
Mask $\boldsymbol{\omega}_i$ to ensure the Wyckoff letters are valid for the given space group g and appears in an alphabetical order.
return $[\boldsymbol{\omega}_1, \boldsymbol{\alpha}_1, \boldsymbol{\chi}_1, \boldsymbol{\upsilon}_1, \boldsymbol{\zeta}_1, \boldsymbol{\omega}_2, \boldsymbol{\alpha}_2, \boldsymbol{\chi}_2, ..., \boldsymbol{\ell}]$

TABLE S1. A table of hyperparameters used in this work.

Hyperparameters	**Value**	**Remarks**
The length of atom sequence including the padding atoms	21	
Number of chemical species	119	'H' to 'Og', plus padding atom
Number of possible Wyckoff letters	28	'a-z'+'A', plus padding atom
Number of modes in von-Mises mixture distribution K_x	16	
Number of modes in lattice Gaussian mixture distribution K_l	16	
Hidden layer dimension for the composite type of the first atom	256	
Transformer number of layers	16	
Transformer number of heads	16	
Transformer key size	64	
Transformer model size d_{model}	32	
Embedding dimension of discrete input	32	
Number of Fourier frequency N_f	5	
Learning rate	0.0001	
Learning rate decay	0.0	
Weight decay	0.0	
Clip grad	1.0	
Batch Size	100	
Optimizer	Adam	
Dropout rate	0.5	
Total number of parameters: 4840295		

2. Sampling algorithm

Algorithm 2 summarizes the sampling method of `CrystalFormer`. It takes 520 seconds to generate a batch size 13,000 crystal samples on a single A100 GPU, which translates to a generation speed 40 millisecond per sample.

Algorithm 2 Sample crystals with `CrystalFormer`

Input: space group number g, a list of chemical elements `element_list`, length n of the atom sequence, sampling temperature T
Output: Wyckoff letters $\boldsymbol{W}$, chemical species $\boldsymbol{A}$, fractional coordinates $\boldsymbol{X}$ of atoms, and lattice parameters $\boldsymbol{L}$ of the unit cell.

Initialize $\boldsymbol{W} = \emptyset$, $\boldsymbol{A} = \emptyset$, $\boldsymbol{X} = \emptyset$
for $i = 1 \ldots, n$ **do**
  # sample Wyckoff letter w
  Get the last $\boldsymbol{\omega}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
  $w \sim \text{Categorical}(\boldsymbol{\omega})^{1/T}$
  $\boldsymbol{W} = [\boldsymbol{W}, w]$
  # sample atom species a
  Get the last $\boldsymbol{\alpha}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
  Mask the logits in $\boldsymbol{\alpha}$ according to `element_list`
  $a \sim \text{Categorical}(\boldsymbol{\alpha})^{1/T}$
  $\boldsymbol{A} = [\boldsymbol{A}, a]$
  # sample fractional coordinate x
  Get the last $\boldsymbol{\chi}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
  $x \sim \text{vonMisesMix}(\boldsymbol{\chi})^{1/T}$
  Project x to Wyckoff positions according to the Wyckoff letter w
  update $\boldsymbol{X}$ with x
  # sample fractional coordinate y
  Get the last $\boldsymbol{\upsilon}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
  ...
  update $\boldsymbol{X}$ with y
  # sample fractional coordinate z
  Get the last $\boldsymbol{\zeta}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
  ...
  update $\boldsymbol{X}$ with z
end for
# sample $\boldsymbol{L}$
Get $\boldsymbol{\ell}$ from `CrystalFormer`$(g, \boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X})$
$\boldsymbol{L} \sim \text{GaussianMix}(\boldsymbol{\ell})^{1/T}$
Symmetrize $\boldsymbol{L}$ according to space group g
return $\boldsymbol{W}$, $\boldsymbol{A}$, $\boldsymbol{X}$, $\boldsymbol{L}$

TABLE S2. Stable crystals sampled in the $Fm\bar{3}m$ space group (No. 225) and the lattice constant of the cubic cell. We list the lattice constant in the bracket if these samples can be found in the MP-20 dataset and are also labelled to be $Fm\bar{3}m$ space group. The materials listed on the left column can be found in the MP-20 dataset. Among them, CeAs is in the test dataset while others are in the training dataset. EuTl, and InNd are in the $Pm\bar{3}m$ space group (No. 221). $InLiY_2$ is in the *Immm* space group (No. 71). The ten materials listed in the right column are not contained in the Materials Project database [61].

In MP			**Not in MP**		
Formula	E_{hull} (eV/atom)	Lattice constant (Å)	Formula	E_{hull} (eV/atom)	Lattice constant (Å)
$PdPtTm_2$	-0.00164	6.90 (6.86)	Ac_2CuRh	-0.14248	7.47
$LiPt_2Zr$	-0.01668	6.43 (6.39)	Ac_2AgSi	-0.04364	7.81
$CdLi_2Pb$	-0.00313	6.82 (6.79)	Ir_2LuPm	-0.0289	6.89
Li_2NdPb	-0.05703	7.05 (7.03)	$InPm_2Tl$	-0.01318	7.73
$BeOs_2Si$	-0.00779	5.77 (5.74)	$PdPm_2Zn$	-0.0843	7.27
CeAs	-0.0073	5.99 (6.09)	$AuInPm_2$	-0.0568	7.70
Xe	-0.00067	7.45 (6.66)	$AcBi_2K$	-0.20215	8.37
$InLiY_2$	-0.17504	7.54	InTb	-0.26278	7.95
InNd	-0.21393	8.35	Ac_2HgIn	-0.40448	8.66
Be_2C	-0.01227	4.33 (4.32)	$CaPaRu_2$	-0.09344	6.80
EuTl	-0.1054	7.90			

Appendix B: Stable crystal samples

Table S2 list generated samples in the $Fm\bar{3}m$ space group (No. 225) with $E_{\text{hull}} < 0$. Three of them (InNd, InTb, and EuTl) actually fall into the $Pm\bar{3}m$ space group (No. 221) due to merging of "4a" and "4b" Wyckoff positions that are occupied by the same elements. Overall, out of 1000 samples generated by `CrystalFormer`, we found 10 stable materials that are not contained in the Materials Project database. Extending the search to more recent material databases, we found match of Ac_2CuRh, Ac_2HgIn, Ac_2AgSi and $AuInPm_2$ in the WBM dataset [93], Ac_2HgIn, Ac_2AgSi, $InPm_2Tl$, $PdPm_2Zn$, Ir_2LuPm and $AuInPm_2$ in the Alexandria dataset [94, 95]. Neverthless, we did not find $AcBi_2K$, InTb, and $CaPaRu_2$ to be present in any of these dataset.

To estimate the energy above hull, the DFT calculations were performed with the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [96] and all-electron projector-augmented wave method [97], as implemented in the VASP code [98]. All parameters of the calculations including settings of PBE functional, Hubbard U corrections, and ferromagnetic initialization are chosen to be consistent with Materials Project by using of `MPRelaxSet` function in pymatgen [60] A double relaxation strategy was employed. The maximum optimization ionic step and the maximum running time were constrained to 150 steps and 20 hours, respectively. All structures containing Yb element are ignored when calculating energy above hull due to they are unavailable from the Materials Project at the time of writing.