Science Bulletin 70 (2025) 3522–3533

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Space group informed transformer for crystalline materials generation

Zhendong Cao [a,b], Xiaoshan Luo [c,d], Jian Lv [c,*], Lei Wang [a,e,*]

[a] Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China
[c] Key Laboratory of Material Simulation Methods and Software of Ministry of Education, College of Physics, Jilin University, Changchun 130012, China
[d] State Key Laboratory of High Pressure and Superhard Materials, College of Physics, Jilin University, Changchun 130012, China
[e] Songshan Lake Materials Laboratory, Dongguan 523808, China

ARTICLE INFO

Article history:
Received 10 February 2025
Received in revised form 5 May 2025
Accepted 12 September 2025
Available online 24 September 2025

Keywords:
Inorganic crystals
Generative model
Autoregressive transformer
Space group symmetry

ABSTRACT

We introduce `CrystalFormer`, a transformer-based autoregressive model specifically designed for space group-controlled generation of crystalline materials. By explicitly incorporating space group symmetry, `CrystalFormer` greatly reduces the effective complexity of crystal space, which is essential 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 coordinates of symmetry-inequivalent atoms in the unit cell. We demonstrate the advantages of `CrystalFormer` in standard tasks such as symmetric structure initialization and element substitution over widely used conventional approaches. Furthermore, we showcase its plug-and-play application to property-guided materials design, highlighting its flexibility. Our analysis reveals that `CrystalFormer` ingests sensible solid-state chemistry knowledge and heuristics by compressing the material dataset, thus enabling systematic exploration of crystalline materials space. The simplicity, generality, and adaptability 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 discovery and design.

1. 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 generate new samples from the learned distribution. In addition, the generation process can also be controlled by conditions 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 pre-trained foundation model for crystalline materials is a key step towards such a lofty goal. However, despite 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 feasible because the amount of high-quality data for crystalline materials is much less 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 [23–26].

The space group symmetry due to the joint outcome of the rotational and translational symmetry in space is arguably the most important inductive bias in the modeling of crystalline materials. There are in total 230 space groups [27] 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 first and the least symmetric space group $P1$ is rare [28], with many instances potentially even being misclassified [29]. Failing to match the space group distribution of nature in machine learning-generated materials is regarded as a matter of serious concern [30].

* Corresponding authors.
E-mail addresses: lvjian@jlu.edu.cn (J. Lv), wanglei@iphy.ac.cn (L. Wang).

https://doi.org/10.1016/j.scib.2025.09.035
2095-9273/

Space group symmetry imposes significant constraints on a crystal. First of all, the space group identifies the crystal system to which a crystal belongs, thereby limiting the permissible values for the lattice parameters that define the length and angles of the crystal's unit cell. Moreover, the symmetry operations associated with a given space group ensure that symmetry equivalent atoms are consistently mapped among themselves in the crystal. This requirement enforces strict conditions regarding the types of chemical elements present, their specific locations within the crystal, and the number of each chemical species in the unit cell. A key concept to express 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. 1a 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. 1a contains the 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' deterministically define the locations of the carbon and calcium atoms within the unit cell. In addition, it follows that $a=b$, and $\alpha=\beta=90^\circ, \gamma=120^\circ$ 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 x-coordinate of oxygen atom 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 Fig. 1 are ubiquitous in crystalline materials. The Wyckoff positions not only specify possible locations of atoms in the unit cell, but 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]. On the other hand, statistical analysis shows that the Wyckoff sequences of known inorganic compounds [31] are far from being exhausted, implying there are statistical correlations to be exploited to compress the materials database.

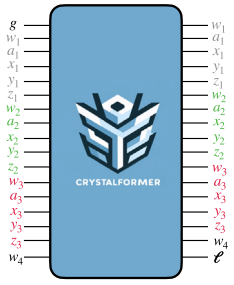
In this paper, we introduce `CrystalFormer`, an autoregressive transformer for generative modeling of crystalline materials. `CrystalFormer` models the joint probability distribution of Wyckoff positions, chemical species, and lattice parameters of crystals with a given space group. By treating the Wyckoff positions as the first class citizen in the model, `CrystalFormer` seamlessly and rigorously integrates the space group symmetry into crystal probabilistic modeling. In the '$P1$ world', one treats crystals as if they were in the first and the least symmetric $P1$ space group. For the estimate, we consider 100 possible chemical elements and 20 atoms in the unit cell with a coordinate grid size of 100 in each direction. The size of the crystalline material space is $N=(10\times 100^3)^{20}\approx 10^{160}$. In the case of utilizing the symmetry of a typical space group, we consider 5 symmetry inequivalent atoms occupying 10 possible Wyckoff positions. The size of the crystalline material space becomes $N=(100\times 10\times 100)^5\approx 10^{25}$. The additional factor of 100 accounts for the remaining degree of freedom for the fractional coordinates and lattice parameters. For

(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)



(c)

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	$\frac{1}{4}$

167-a-C-0-0-1/4-b-Ca-0-0-0-e-O-0.257-0-1/4

(d)

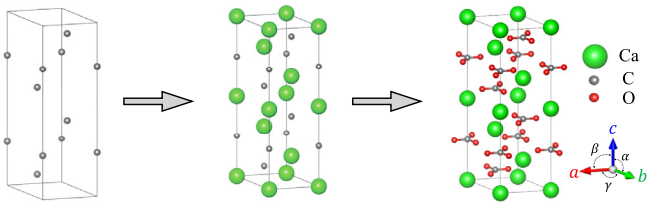


Fig. 1. (a) The Wyckoff positions of the $R\bar{3}c$ space group (No. 167). We highlight the occupied Wyckoff positions of calcite $CaCO_3$ crystal which belongs to this space group. Carbon, calcium, and oxygen atoms occupy the '6a', '6b', and '18e' positions, respectively. (b) The `CrystalFormer` is 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 symmetry inequivalent atom, and finally, the lattice parametrized by ℓ sequentially. (c) The crystal data of $CaCO_3$ are summarized in a table. In the table, the x-coordinate of oxygen atom $x_3=0.257$ 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 letters. The string below the table shows the sequential representation of the $CaCO_3$ crystal with space group, Wyckoff letter, and atom species as the input to the `CrystalFormer` model. (d) Autoregressive generation of the 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, atom species, and the fractional coordinates if they are still unspecified.

alternate estimates of the materials space in the context of crystal structure prediction, see Refs. [32,33]. As analyzed above, explicit modeling of the Wyckoff positions greatly reduces the space of crystalline materials. The space group-informed transformer exploits this fundamental inductive bias to greatly simplify the learning and generation of crystals.

2. Method

We will first introduce the `CrystalFormer` model, then reveal the chemical intuition encoded in the trained model by inspecting generated crystal samples. These inspections also build up understandings of the strength of the model.

2.1. `CrystalFormer`

We will introduce the design, training, and sampling of the `CrystalFormer` model.

2.1.1. Model

To exploit the space group symmetry of the crystal, we focus on the Wyckoff positions of symmetry-inequivalent atoms. Wyckoff letters follow the alphabetical order, where 'a' stands for the positions with the highest order of site 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 of the Wyckoff positions. In cases where the atom positions are not 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 sites with larger multiplicities.

With these considerations, we define a crystal data as $\mathcal{C} = \{\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{L}\}$. 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. 1b 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 crystal.

The central quantity to focus on is the conditional probability of a crystal $\mathcal{C}$ given the space group number $g \in [1, 230]$: $p(\mathcal{C}|g)$. 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 of crystal structure prediction and material design, the space group can either be considered separately as a control variable or predicted based on material composition [34–37].

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

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

At first sight, it may appear unnatural to employ an autoregressive model for crystals since there seems to be 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. Following this key observation, we represent crystal data as sequences of space groups, Wyckoff letters, chemical species, and fractional coordinates of each symmetrically inequivalent atom. Together with the information lattice parameters, such sequence fully characterizes the compositional and structural information of crystalline material. Since statistical analysis reveals that anions are in less symmetric positions than cations for inorganic crystals [28], one would expect that anion atoms will typically appear after cation atoms in such a sequence. For example, $CaCO_3$ is represented as a string '167-a-C-0-0-1/4-b-Ca-0-0-0-e-O-0.257-0-1/4'. Autoregressive sampling of such a string means the model generates the crystal by placing the atoms sequentially into the unit cell, starting from the special position with high site symmetry to the general position with the lowest site symmetry, see Fig. 1d.

We model the conditional probability of the Wyckoff letters $\boldsymbol{W}$ and chemical species $\boldsymbol{A}$ as categorical distributions. On the other hand, we model the conditional probability of the factional coordinates $\boldsymbol{X}$ as a mixture of von Mises distribution for continuous periodic variables. For Wyckoff positions with multiplicities greater than one, we only consider the first of fractional coordinates that appear in the international tables for crystallography [38]. Lastly, we model the conditioned distribution of lattice parameters as a Gaussian mixture model.

We build `CrystalFormer`, an autoregressive transformer [39] to model the space group conditioned-probability distribution of crystalline materials Eq. (1). The space group number g is the first input to `CrystalFormer`. The remaining inputs are the Wyckoff letter, chemical species, and fractional coordinates of each atom. One can go through the table of Fig. 1b 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 also concatenate the vector embedding of g to all other inputs since it is the key control variable for the crystal generation. Moreover, 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 [13,40]. We pad the atom sequence up to a maximum length and treat the output as parameters of the conditional probability distribution Eq. (1), see Fig. 1b. 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 its phase space. First, the Wyckoff letters should be valid for the given space group. For example, for the space group $R\bar{3}c$ (No. 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. We follow the convention that capital letters appear after lower case letters. This handles the edge case of the *Pmmm* space group (No. 47) whose Wyckoff letters used up 26 lowercase letters and reached 'A' for the generic position. In addition, we use the letter 'X' to indicate the Wyckoff position of padding atoms that appear at the end of the sequence, see e.g., w_4 of Fig. 1b. Lastly, the Wyckoff positions with no free fractional coordinates (such as 'a', 'b', and 'd' positions in the $R\bar{3}c$ space group) can only be occupied once. Those constraints are implemented by setting the logit biases of Wyckoff letters to mask out invalid sequences [41].[1]

[1] OpenAI, Using logit bias to alter token probability with the openai api. OpenAI Help Center

The design of CrystalFormer focuses mostly on the space group symmetries which we believe to be the most important inductive bias 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 over atom of the same type 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 [42] 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 in CrystalFormer since they are treated as periodic variables following the von Mises distribution.

2.1.2. Training

The CrystalFormer is trained by minimizing the negative log-likelihood

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

where the structures $\mathcal{C}$ and the corrosponding space group g of crystals are sampled from the training dataset. Writing out $p(\mathcal{C}|g)$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 $\boldsymbol{L}$. For the continuous variables $\boldsymbol{X},\boldsymbol{L}$ in the objective function, 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., $90^\circ, 120^\circ$) which were already fixed by the chosen space group and Wyckoff letter will not not contribute to the loss function.

In the present work, we train the CrystalFormer using the MP-20 dataset [11]. MP-20 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 imbalanced distribution of crystals over space groups in nature [28]. In fact, there is no training data in 61 out of 230 space groups as shown in Fig. S2 (online). 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, see Section B in the Supplementary material. Note that the CrystalFormer can generate reasonable samples even for those space groups without any training data. This because the model can exploit knowledge learned from other space groups to place suitable atoms in the Wyckoff positions due to weight sharing. Moreover, since the sampling process makes use the of Wyckoff position table. The three dimensional coordinates of atoms are not completely random even for unseen space groups. Fig. 2 shows a breakup of the learning curves for the Wyckoff position, chemical species, fractional coordinates, and lattice parameters. We select the model checkpoint with the lowest total validation loss to generate crystal samples.

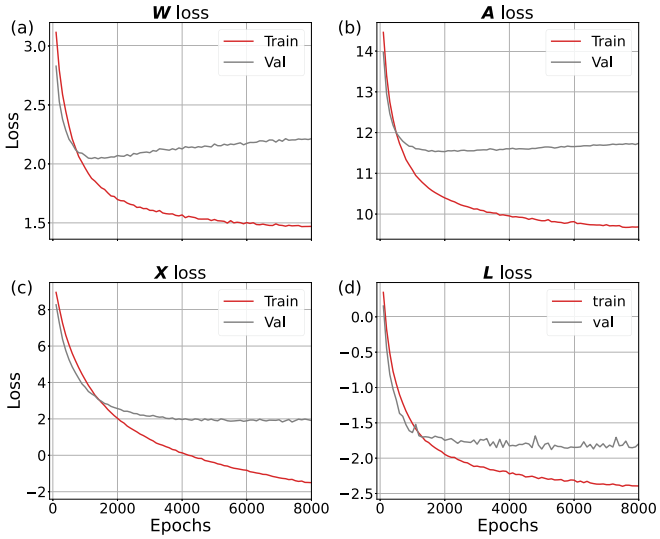


Fig. 2. Break up of the training and validation losses for (a) Wyckoff letters, (b) chemical species, (c) fractional coordinates, and (d) lattice parameters over training epochs.

2.1.3. Sampling

To sample crystals from the CrystalFormer, one needs to specify a space group number and a list 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 rectifies it and 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 inequivalent 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 $p(\mathcal{C}|g)^{1/T}$. With $T<1$ we will draw samples from a sharper distribution, while $T>1$ gives more diversity in the generated samples. In the present paper, we will generate crystals using temperature $T=1$ unless mentioned explicitly.

Besides autoregressive sampling, one can also perform Markov chain Monte Carlo (MCMC) sampling based on the likelihood Eq. (1) of the CrystalFormer. MCMC sampling can walk through the crystalline materials space starting from an existing crystal structure. At each step of the random walk, one proposes a configuratoin update in terms of element substitution, atom position shift, or lattice deformation to change the crystal from $\mathcal{C}$ to $\mathcal{C}'$, then accepts or rejects the proposal according to the model probability following the Metropolis acceptance rule $\min\left[1,\frac{p(\mathcal{C}'|g)}{p(\mathcal{C}|g)}\right]$. MCMC sampling is particularly useful for incorporating additional constraints or guidance in the sampling procedure. Moreover, 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.

2.2. CrystalFormer learns chemical intuition by compressing materials database

Nature favors symmetrical crystal structures. Crystallographic space groups quantify this inductive bias of nature, thereby significantly simplifying the spaces of crystal materials. In light of the

space group symmetries, crystals also have an unexpected yet natural sequential and discrete representation, which derives from two tables in nature: the periodic table of elements determined by quantum mechanics and the table of Wyckoff positions of the 230 space groups determined by group theory. To construct a certain crystal, we only need to select atoms from the periodic table and place them sequentially into the Wyckoff positions in the unit cell. In this crystal language, the 'word order' is determined by the alphabetical order of Wyckoff letters, the 'grammar' corresponds to the solid-state chemistry rules, and the 'synonyms' represent interchangeable elements (Section 2.2.1), the 'sentence length' correspond to atom number in the unit cell (Section 2.2.2), and the 'idioms' correspond to common chemical coordination (Section 2.2.3).

`CrystalFormer` employs an autoregressive transformer to learn the crystal language, thereby exploring yet-to-be-discovered crystalline materials. It compresses and internalizes the crystal materials database, expressing solid-state chemical knowledge through neural network parameters; reflecting the associative ability of material space through neural network activations; and describing chemical intuition through the model probability (Section 2.2.4). Similar to generative models used for generating text, images, and videos, `CrystalFormer` can directly generate 'realistic' crystal materials. However, rather than worrying about the fake contents of AI-generated media, these AI-generated crystal materials could potentially be synthesized and be useful to human civilization.

Next, we will inspect the learned features and sample statistics of the model to build up an understanding of the `CrystalFormer`. We carry out inspectations for a few selected space groups. The findings are neverthelss general. These findings provide understandings and confidence of the model, therefore direct us to the suitable applications of `CrystalFormer`.

2.2.1. Atom embeddings and chemical similarity

Fig. 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 sees the chemical similarity within groups of elements show up as off-diagonal red 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). The features shown in Fig. 3 are strikingly similar to the similarity map constructed purposely based on substitution pattern [43,44] which was later used for substitution-based material discoveries [5,45]. In the context of language modeling, the chemical similarities correspond to synonyms of chemical species tokens. Having the ability to learn chemical similarities from data [19,43,44,46–50] is an encouraging signal that the model is picking up atomic physics for generating reasonable crystal structures with maximum likelihood based training.

2.2.2. Atom number distributions

The number of atoms in the unit cell corresponds to the length of non-padding atoms in the crystal string representation, which is captured well by `CrystalFormer`. Fig. 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. Incorporating Wyckoff positions information into the `CrystalFormer` model architecture removes the necessity of querying the training data to find out the number of atoms for a targeted space group [16] during generation.

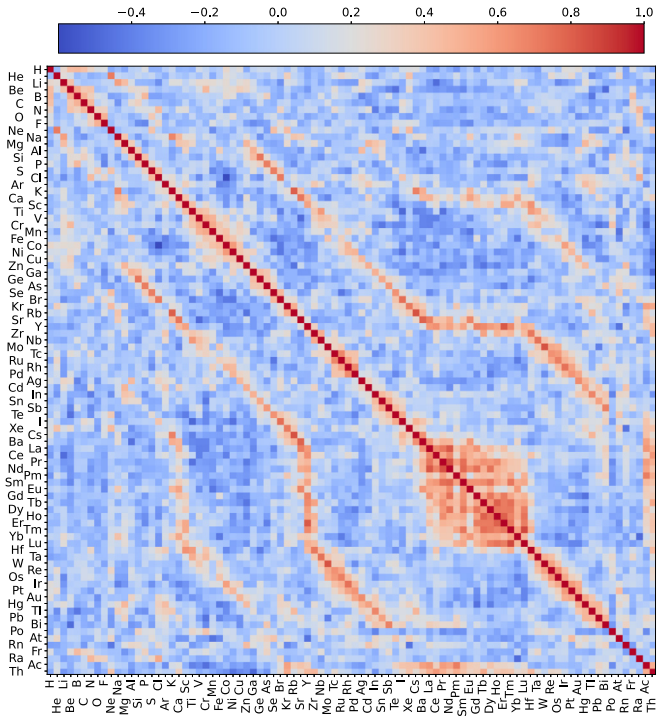


Fig. 3. The cosine similarity matrix for the chemical species based on the learned vector embeddings. The reddish color suggests similar chemical elements in the crystal environment.

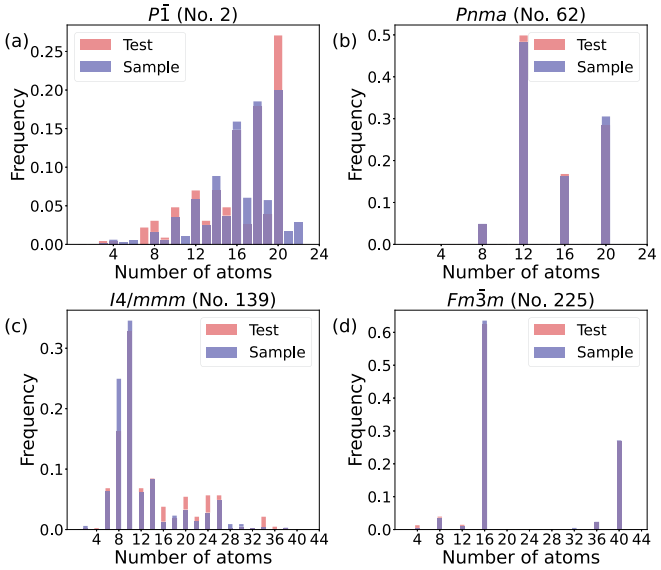


Fig. 4. The histogram for total number of atoms in the unit cell for (a) $P\bar{1}$, (b) *Pnma*, (c) $I4/mmm$, (d) $Fm\bar{3}m$ space groups in the test dataset and in the generated samples.

Recently, Ref. [51] reports an abundance of inorganic compounds whose primitive unit cell contains a number of atoms that is a multiple of four. There are different ways to reason about the observed 'rule of four' depending on one's view of how a crystal is formed. For example, one can often break inorganic solids into polyhedra as building blocks. Otherwise, Ref. [52] considers the

most probable values of the number of atoms in a formula unit and the number of formula units per primitive cell. In line with the discussion here, the 'rule of four' is the combination of three factors: (1) the distribution of crystalline materials among space groups [28]; (2) the distribution of atoms in Wyckoff positions [31] of a given space group; and (3) the multiplicities of Wyckoff positions and multiplicities of conventional versus primitive cells. The first two are statistical rules determined by the inter-atomic interactions while the third one is a mathematical fact of space group theory. In the end, the point we want to make is that the design of `CrystalFormer` and its associated crystal representation allow it to learn the 'rule of four' and many other to-be-discovered 'rules', which manifest themselves as marginal statistics of learned probability distribution. Most importantly, `CrystalFormer` will utilize these 'empirical rules' when generating novel yet reasonable crystal samples.

2.2.3. Wyckoff-Atom gram

Fig. 5 shows heat maps of Wyckoff positions and chemical species for the $Fm\bar{3}m$ space group (No. 225). The heat map is analogous to bigram frequency statistics in language modeling. In the present context, it reveals interesting solid-state chemistry knowledge related to where each atom tends to appear in a unit cell. First of all, one sees that most atoms occupy special Wyckoff positions (Wyckoff letters at the beginning of the alphabet) with higher site symmetries. The distribution of generated data is in agreement with test data and recent statistics [31]. Moreover, there are vertical blanks at the locations of inert elements (He, Ne, Ar, ...) as they are rare in crystalline materials. Lastly, one sees that oxygen and halogen elements (F, Cl, Br, I) appear quite often in the Wyckoff position '24e', which means these high electronegative elements form polyhedra enviorment for other atoms [28]. Overall, we see the `CrystalFormer` has learned these key motifs for generating crystalline materials. On the other hand, one also observes 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.

Along the same line of thoughts, coordination polyhedra [53] and lattice structure [54] manifest themselves as higher-order n-gram correlations of Wyckoff position and atom species in the crystal language, which will be captured by the `CrystalFormer`. There has been a long history of mining empirical chemistry rules encoded in materials data and then using them to instruct the search of crystal structures [44,55–59]. Our analysis shows that `CrystalFormer` ingests chemical intuition, be it speakable or unspeakable, in the training data for generating new materials.

2.2.4. Crystal likelihoods

`CrystalFormer` compresses chemistry knowledge stored in the material dataset into its parameters. In addition to generating crystal samples, `CrystalFormer` can also compute the likelihoods of crystals via Eq. (1). Therefore, it is also possible to employ `CrystalFormer` in likelihood-based Monte Carlo search besides sampling crystals directly.

Fig. 6 shows the agreement of the likelihoods of generated samples and samples in the test dataset. We also visualize structures of a few generated samples which are deemed to be very likely, typical, and unlikely according to their likelihood values. We have checked that likelihood is related to the energy of the crystal by locally perturbing the fractional coordinates and lattice parameters. However, we did not observe a correlation between the likelihood of these crystals and their energies on a global scale. We envision the landscape of likelihoods is much less rough compared to the potential energy surface of crystalline materials. Intuitively, it means that the `Crystalformer` compresses the materials space into a more compact space without many holes that correspond to infeasible high energy states. Therefore, likelihood-based exploration of the crystal space discussed in Section 2.1.3 can be more efficient compared to traditional sampling approaches based on the Boltzmann distribution based on physical energy functions.

3. Results

We now move on to the practical applications of `CrystalFormer` to materials discovery and design. Compared to many existing materials generation models, `CrystalFormer` offers precise control over space group symmetry and enables efficient computation of model likelihood. These unique features open a wide range of possibilities for integrating it with existing computational software and machine-learning models in a flexible way as we demonstrate below. For these applications, we have excluded radioactive elements from the samples [30].

3.1. Symmetry-conditioned random structure initialization

Crystal structure prediction has long been the dream of solid-state chemistry and computational material science researchers [60]. Typical crystal structure prediction workflow consists of two steps. First, one randomly initializes a batch of diverse crystal structures as candidates. Second, one optimizes the crystal structures via local and global optimization strategies. Utilizing space group symmetries plays a crucial role in both steps, as symmetry enlarges the span of the energy distribution [61–63] and reduces the search space.

It is a common practice for crystal structure prediction software [62–67] and structure search [68–70] to exploit space group symmetry in the crystal structure initialization. However, such an initialization approach faces combinatorial difficulty as the number of chemical species and atoms in the unit cell grows. The `CrystalFormer` is ready to act as a drop-in replacement of random structure initialization for crystal structure prediction. In this way, one bypasses the curse-of-dimensionality of exact enumeration [62] with a data-driven probabilistic approach. Moreover, the ability of `CrystalFormer` to generate diverse and near-stable structures can greatly reduce the computational costs of downstream optimizations.

We select seven space groups $P\bar{1}$ (No. 2), $C2/m$ (No. 12), $Pnma$ (No. 62), $I4/mmm$ (No. 139), $R\bar{3}m$ (No. 166), $P6_3/mmc$ (No. 194), and $Fm\bar{3}m$ (No. 225) as representatives of the seven crystal systems. We randomly generate 100 crystals for each space group using `CrystalFormer`. On the other hand, we employ PyXtal [63] to generate crystal samples with the same stoichiometry in the same space groups. We then carry out structure relaxation using density functional (DFT) calculations.

Fig. 7a–g shows the average energy difference to the energy of final structures versus DFT relaxation steps. We neglected the structures whose energy changes and energy change intervals per step during relaxations exceeded 10 eV/atom to eliminate the impact of erroneous steps. One sees that `CrystalFormer` samples generally reach lower energies in fewer relaxation steps. This is especially true for space groups with higher symmetries. The ability to initialize diverse and high-quality crystal structures enables one to discover more stable materials faster. Fig. 7h shows the histogram of energy above the convex hull constructed by the Materials Project database. The dashed line denotes the criterion $E_{\text{hull}} < 0.1$ eV/atom [71] for selecting stable materials. Among these candidates, we found 34 and 12 relaxed structures with `Crys-`

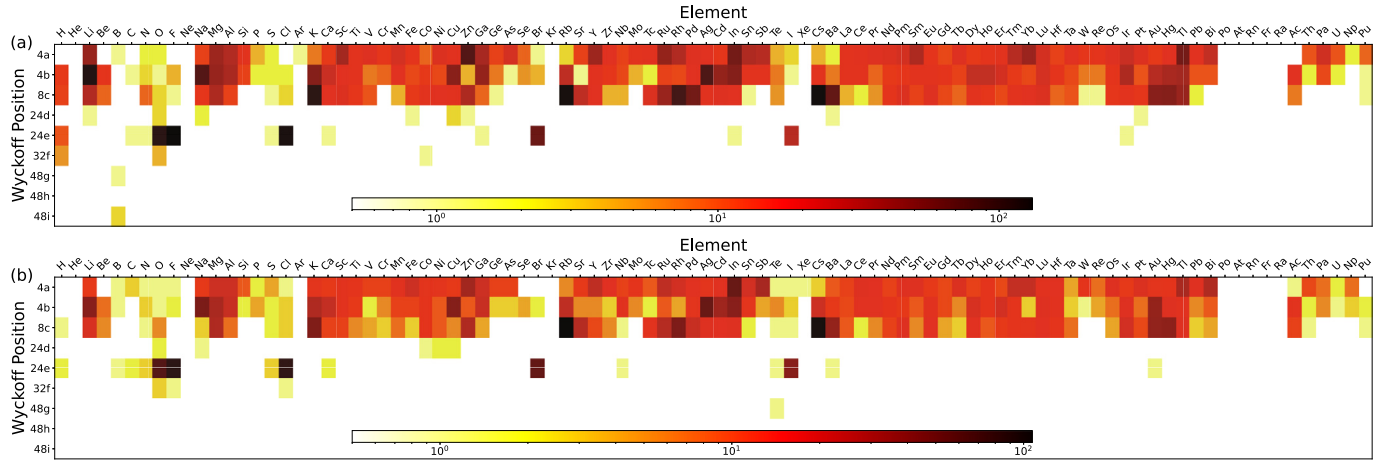


Fig. 5. The heat map for Wyckoff positions and atom species of (a) the test dataset and (b) generated samples for the $Fm\bar{3}m$ space group (No. 225). It is an analog of bigram frequency statistics of language modeling, which shows the prefereed Wyckoff positions of different chemical species.

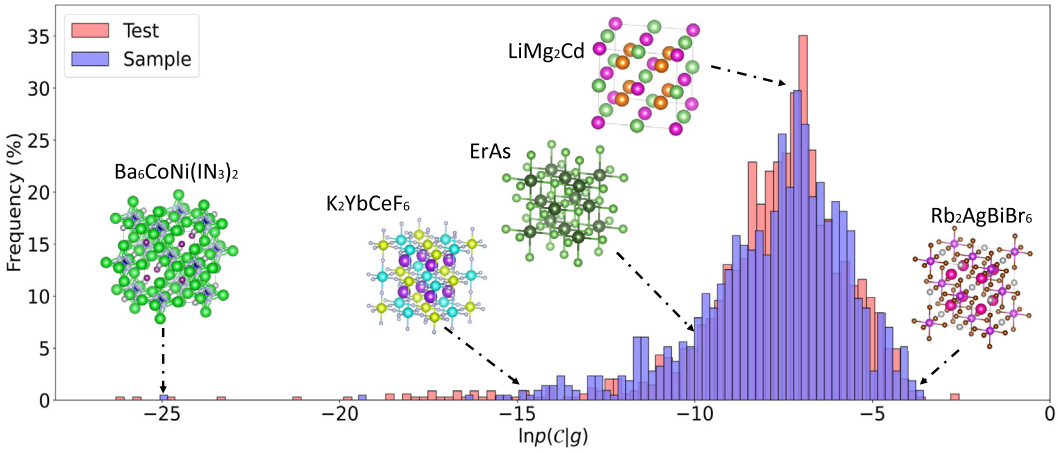


Fig. 6. The histogram of log-likelihoods of 1000 samples in the $Fm\bar{3}m$ (No. 225) space group and the test dataset. The insets visualize the crystal structure of a few generated samples. $Rb_2AgBiBr_6$ and $LiMg_2Cd$ are in the training dataset. ErAs is in the validation dataset. K_2YbCeF_6 and $Ba_6CoNi(IN_3)_2$ are not in the MP-20 or Materials Project database.
.

talFormer and PyXtal initializations that are not contained in the MP-20 dataset. We summarize them in Tables S3 and S4 of Section C in the Supplementary materials.

Table 1 lists detailed statistics of structure-relaxed samples in seven representative space groups. Overall, we find that the CrystalFormer generated structures are of higher quality, especially for those space groups with higher symmetry. This observation is supported by the fact that the DFT relaxation often retains the space group symmetry. The root mean squared displacement (RMSD) [16] computed for these converged structures demonstrate CrystalFormer's superior performance over PyXtal across all 7 space groups. The average energy above the convex hull also confirms the samples generated by CrystalFormer are indeed much closer to the DFT local minimum than PyXtal initialization. CrystalFormer attains superior performance in the high-symmetry space groups compared to the RMSD of 0.11 Å reported in Ref. [16] for the MatterGen model trained on the MP-20 dataset with no control on the space-group symmetry.

3.2. Structure-conditioned element substitution

Mutation of known crystals is a prominent approach to materials discovery. For example, one can employ a machine-learned force field to relax crystal structures [5,72–74] 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 for the limitation of energy-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.

Many crystal structures can be traced back to a few simple, highly symmetrical types. Numerous crystals share the same structural prototype but differ in composition, such as perovskite (ABX_3), spinel (AB_2X_4), fluorite (AX_2), and so on. Fig. 8a shows double perovskite crystal structures $A_2BB'X_6$ which belong to the $Fm\bar{3}m$ (No. 225) space group. There are hundreds of known double perovskites with significant interests in their semiconducting, ferroelectric, thermoelectric, and superconducting properties [75]. Finding more stable materials with this structure prototype using brute force enumeration and high-throughput calculation is a computationally demanding task [76]. We will generate new double perovskites with CrystalFormer and demonstrate its advantage of over standard element substitution methods.

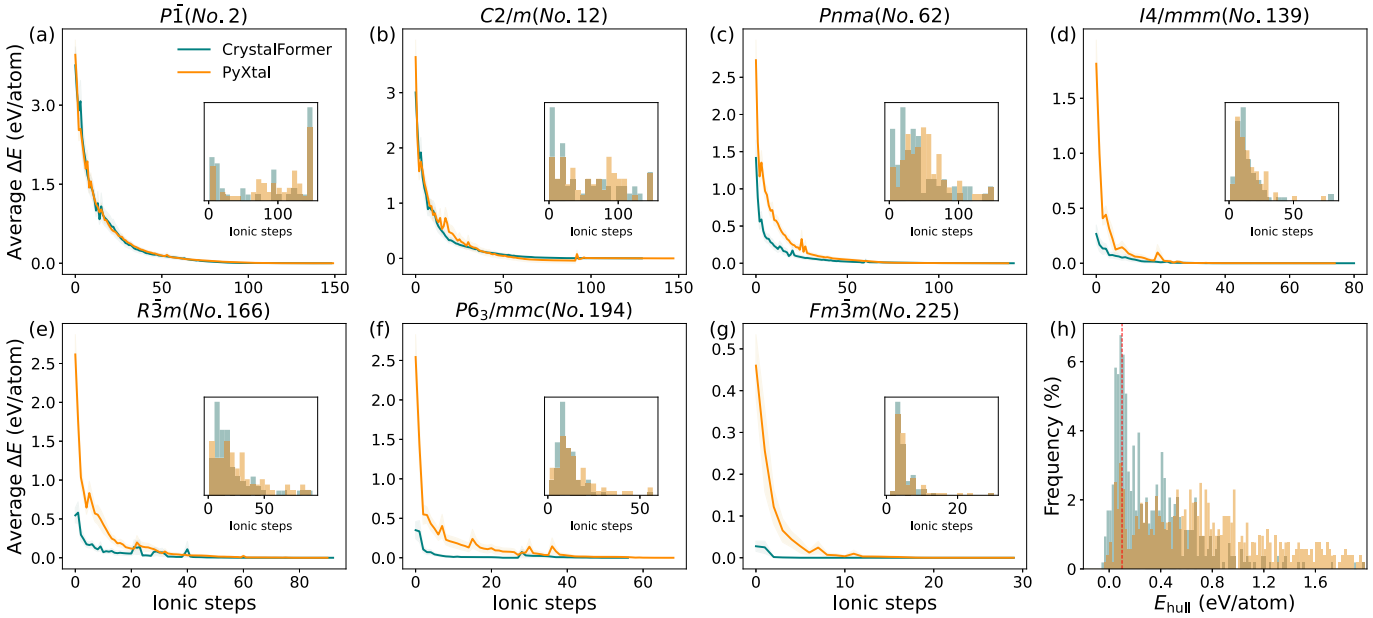


Fig. 7. (a)–(g) Average energy difference versus relaxation steps for seven representative space groups. The insets show the distribution of ionic steps. (h) The histograms of energy above the convex hull for relaxed crystal structures. The dashed line indicates the criterion for selecting candidates for stable materials listed in Section C in the Supplementary materials since materials with $E_{hull} < 0.1$ eV/atom are usually metastable and have the potential to be synthesized [71].

Table 1
For each space group we randomly generate 100 crystal structures with the same composition using `CrystalFormer` and PyXtal. We carry out energy relaxation using DFT calculations and report the number of converged samples, the number of structures that maintain the original space group symmetry, the average RMSD between generated and relaxed structures, and the averaged energy above the convex hull.

Space group	Crystal system	Converged structures ↑		Retain symmetry ↑		RMSD[1] (Å) ↓		E_{hull}[1] (eV/atom) ↓	
		`CrystalFormer`	PyXtal	`CrystalFormer`	PyXtal	`CrystalFormer`	PyXtal	`CrystalFormer`	PyXtal
$P\bar{1}$ (No. 2)	Triclinic	46	**67**	45	**67**	**1.181**	1.259	1.034	**0.913**
$C2/m$ (No. 12)	Monoclinic	55	**72**	53	**67**	**1.051**	1.227	**1.233**	1.660
Pnma (No. 62)	Orthorhombic	77	**83**	**76**	66	**0.594**	1.092	**0.313**	1.633
$I4/mmm$ (No. 139)	Tetragonal	**91**	81	**88**	63	**0.140**	0.675	**0.240**	1.100
$R\bar{3}m$ (No. 166)	Trigonal	**83**	74	**80**	71	**0.294**	0.919	**0.352**	2.489
$P6_3/mmc$ (No. 194)	Hexagonal	**97**	77	**96**	60	**0.086**	0.545	**0.324**	4.100
$Fm\bar{3}m$ (No. 225)	Cubic	**98**	96	**95**	92	**0.012**	0.033	**0.214**	0.483

[1] Calculated on the converged structures.

Fig. 8a shows the string representation of double perovskites. To generate candidates of double perovskites, we use `CrystalFormer` to carry out string in-filling tasks. Since the autoregressive sampling of the atoms is insufficient to take into account non-causal information in the sequence, we employ MCMC to sweep through the sequence and update chemical species and fractional coordinates [77]. The acceptance rate for these MCMC updates makes use of the marginalized probability for elements and fractional coordinates as the lattice parameter that appears at the end of the sequence can be integrated. Only after the MCMC sampling has been thermalized, we sample the lattice parameters autoregressive to account for the adjustment of the unit cell for given atoms and occupations. We use `CrystalFormer` to generate 100 candidates as the initial DFT relaxation.

As a comparison, we also employ the `SubstitutionPredictorTransformation` function [43] implemented in pymatgen [78] to perform element substitution for the crystals with double perovskite structures in the training dataset. The substitution probabilities come from data-mining of ICSD dataset [43]. After the substitution, we use `DLSVolumePredictor` [79] function of pymatgen to predict the volume of the structure. This lattice scaling scheme relies on data-mined bond lengths to predict the crystal volume of a given structure. To collect 100 candidates in the ionic substitution approach we have set the probability threshold of `SubstitutionPredictorTransformation` to 0.01, which is smaller than the typical values adopted in Ref. [45].

The RMSD computed for the DFT-relaxed structures is 0.084 and 0.031 Å for `CrystalFormer` and ionic substitution [43], respectively. Moreover, Fig. 8b shows the histogram of energy above the convex hull of the Materials Project database. Overall, `CrystalFormer` and ionic substitution [43] found 9 and 3 double perovskites with $E_{hull} < 0.1$eV/atom which are not contained in the MP-20 dataset, details in Section D of Supplementary materials. The superior performance of `CrystalFormer`-guided MCMC is understandable since its likelihood takes into account the context of space group and atomic environment rather than marginal two-body correlation [43] in ionic substitution. The ionic substitution approach also shows two limitations in practical applications. First, some of the ions in the compound can not be substituted as they are missing in the probability table. Second, the approach relies on the calculability of the elements' valence states which are not always well defined.

As a final remark, although the discussion here focuses on generating crystals with given prototype structures, the generation of crystals with a given crystal lattice [80] is also feasible with `CrystalFormer`. This is because the crystal lattice can be straightfor-

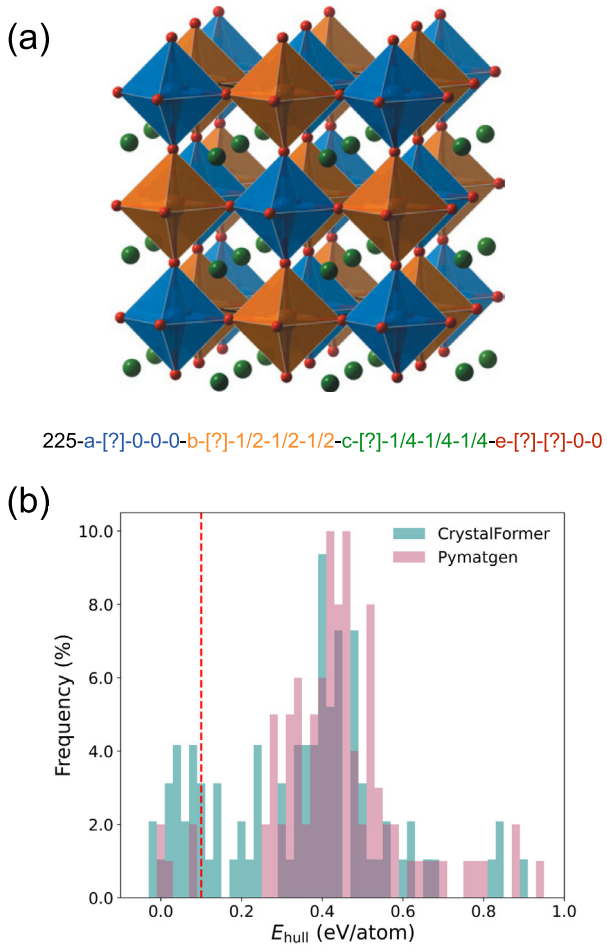


Fig. 8. (a) Double perovskites crystal structure. The crystal string representation of double perovskites with blank spaces for chemical elements and the x-coordinate of the atom resides in the 'e' position. `CrystalFormer` generates crystals with double perovskite structures via sequence infilling. (b) The histograms of energy above the convex hull for the relaxed crystal structures. The dashed line indicates the criterion for selecting candidates for the stable materials.

wardly expressed as constraints on the space group and occupied Wyckoff letters [54].

3.3. Plug-and-play materials design

Finally, we demonstrate `CrystalFormer`'s ability to aid property-guided exploration of crystalline materials in a versatile and flexible manner. The trained `CrystalFormer` captures the space group conditioned crystal probability $p(\mathcal{C}|g)$, which we treat as a prior probability for stable crystals. By combining it with a crystal property prediction model that provides the forward likelihood probability $p(y|\mathcal{C})$, one can carry out property-guided materials generation in a plug-and-play manner. According to Bayes' rule, the posterior for crystals given property y reads

$$p(\mathcal{C}|g,y) \propto p(y|\mathcal{C})p(\mathcal{C}|g). \tag{3}$$

By sampling from this posterior distribution, one can generate crystal samples with property guidance. Since the posterior probability Eq. (3) typically does not process autoregressive property with respect to $\mathcal{C}$, we carry out MCMC sampling to sample from the posterior distribution [81]. The plug-and-play feature makes designing crystalline materials in this way particularly appealing because it is possible to apply multiple conditions by simply adding log-likelihoods from multiple predictors. The framework applies to the inverse problem of solving cyrystal structures based on experimently observed diffraction spectra equally well [82,83], where the goal is to simultaneously optimizing the matching probability to experimental observation and stability of the crystal.

Any property prediction model can be used in conjunction with `CrystalFormer` for property-guided material generation. We utilize two pre-trained MEGNet [48,84] models to predict the band gap and formation energy, using the output of these two property prediction models as the forward probability $p(y|\mathcal{C})$ of Eq. (3). More details are in Section E of the Supplementary materials.

Fig. 9a demonstrates the controlled generation of materials with target band gap at $E_g = 2$ eV and the formation of energy $E_{form} = -3$ eV/atom crystals [85,86]. The conditional probability Eq. (3) contains both the model likelihood and the property regression MAE. Therefore, the generated samples will strike a balance between the two. To draw samples from the conditional probability distribution, we randomly generate a batch of 1000 crystal samples and sweep through the crystal sequence to update the atom species, fractional coordinates, and lattice parameters. For simplicity of the Wyckoff sequence is kept unchanged in the MCMC sampling. Achieving the desired properties via Monte Carlo update of chemical species can be regarded as a systematic data-informed way of carrying out cation-transmutation for materials inverse design [87]. After reaching equilibrium, the histogram of band gap and formation energy is centered around the target values, which is shifted significantly away from the value of unconditionally generated samples. On the other hand, the likelihood shown in Fig. 9b indicates these conditional-generated crystals are not typical samples with respect to the unconditional distribution. Nevertheless, they are still probable samples according to the crystal prior given by the `CrystalFormer`. Note that controlling the distribution of formation energy can significantly impact the distribution of the energy above the convex hull due to the correlation between these two energies. To assess the stability of the conditionally generated samples, we first use M3GNet [72] to filter out the unstable materials, followed by DFT verification on the remaining subset. The conditionally generated samples with $E_{hull} < 0.1$ eV/atom that are not included in the MP-20 dataset are listed in Table S7 in Section E of the Supplementary material. We observed that some materials meet the requirements of property predictors, but not the DFT calculations. This is due to due to the errors in the property prediction models, which could be further improved by employing a more robust crystal property prediction models. Given a myriad of materials property prediction models developed over the years and the inconvenience of re-training or fine-tuning the foundational generative model [16], we envision the plug-and-play generation approach demonstrated here to be a scalable way for materials design. We have exposed an interface of `CrystalFormer` in our code repository for users to plug in arbitrary conditioners for guided materials generation.

4. Discussion and conclusion

4.1. Related works

Crystal generative models have been explored using variational autoencoder [23,88], generative adversarial networks [12,89], normalizing flows [90–93], diffusion models [11–16,21,22,26], GFlowNet [24,25], and autoregressive models [17–20,94–96]. In these autoregressive models, one either uses atomistic features [17,94–96] or uses pure text descriptions [18–20]. Nevertheless, with the introduction of specialized tokens for crystals, the boundary between the two is blurred.

The `CrystalFormer` is most closely related to the autoregressive generative model originally designed for molecules [94–96]. However, instead of predicting the relative distances of atoms,

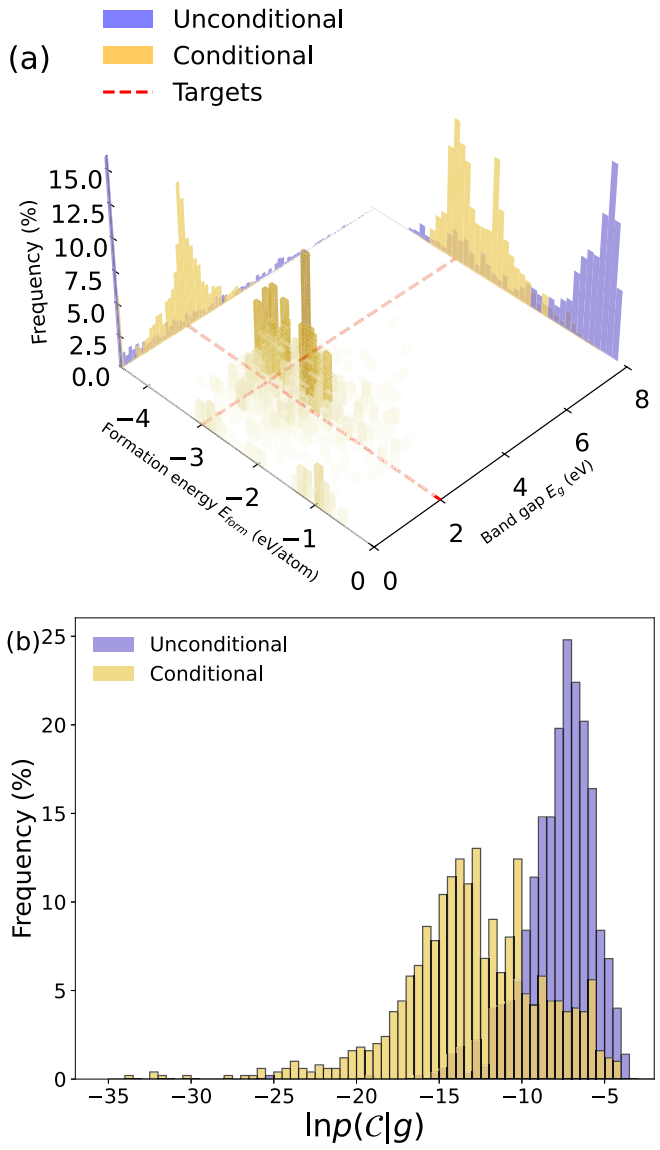


Fig. 9. (a) The histogram of band gap and formation energy predicted by MEGNet models for crystal samples generated in the $Fm\bar{3}m$ space group (No. 225). The dashed red lines in the plane indicate target values. The marginals on the side show the shift of the property distributions with respect to the unconditionally generated samples. Note that we scale the 3d histograms for better visualization. (b) The likelihoods of conditioned generated samples compared to the unconditional samples.

we predict the Wyckoff positions of symmetry-inequivalent 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 spatial symmetries. On the other hand, compared to Ref. [19] which treats text descriptions of crystals using autoregressive nature language model, `CrystalFormer` speaks native crystal language: it deals with a more concise and essential atomistic representation of crystals, which leads to a smaller model size and faster sampling speed. Fast generation speed is not only a welcoming feature but also will be crucial for further exploration of materials space based on combinations of probabilistic generation and post-selection, Monte Carlo sampling, backtracking, and searching techniques [97]. More importantly, by baking in the space group symmetry in the model rather than learning them as statistical correlation from texts [18,20], `CrystalFormer` guarantees space group constraints and cherishes the precious data and computing time. In this sense, the present work employs rigorous mathematical (as opposed to vague 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 [46,98]. Incorporating space group information in the encoder-only transformer models may also enhance their property prediction performance [99–101] as suggested by Ref. [102].

4.2. Outlook

Precisely controlling the space group in the generative model of crystalline materials not only greatly simplifies the task but also is a highly desired feature for materials discovery and design. `CrystalFormer` integrates exact symmetry principles from math and empirical chemical intuitions from data into one unified framework. Probabilistic generative modeling of crystalline materials using `CrystalFormer` opens the way to many future innovations in materials design and discovery.

Note that the MP-20 dataset has by no means exhausted all available crystalline material [16,19]. 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. A later version of `CrystalFormer` which is trained on curated Alex-20 dataset [103] has shown significantly improved performance [104]. In particular, extending the dataset to include both inorganic and organic crystals [105] may be beneficial as it improves the data coverage of low symmetric space groups. The transformer-based generative model is ready to be scaled up to work with much larger and more diverse training data, in the same fashion as large language models [106]. 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 [107] as it has also been showing up in other contexts of atomistic modeling [108].

The model architecture and sampling strategy are both open to further refinement 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 Euclidean normalizer [109] and group-subgroup relation [110] in the model architecture or training procedure. Second, it is worth exploring using `CrystalFormer` as the base distribution in the flow model and employs symmetry-persevering transportation to further adjust the atoms coordinates and unit cells [10,26], which mimics a symmetry-constrained relaxation process [111]. Lastly, it may be worth employing more advanced constrained and guided sequence generation methods [112–115] for more flexible control on the elements, structure, or stoichiometry of generated materials.

Conditioned materials generation depending on properties [16,21,22,96] and experimental measurements [116] are highly desired features of materials generative model. Although it is straightforward to extend `CrystalFormer` (e.g., extend the space group embedding or employ the encoder-decoder transformer architecture [39]) to incorporate these conditions, we are particularly excited about the plug-and-play routine demonstrated in Section 3.3. Along this line, we envision an ecosystem [117] where the foundational generative model for $p(\mathcal{C}|g)$ and more specialized discriminative models for materials properties $p(y|\mathcal{C})$ are developed separately but brought together via the Bayes rule.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (T2225018, 92270107, 12188101, T2121001, and 12034009), the National Key Projects for Research and Development of China (2021YFA1400400), and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0500000 and XDB30000000). We thank Han Wang, Lin Yao, Linfeng Zhang, Chen Fang, Yanchao Wang, Zhenyu Wang, Qi Yang, Shigang Ou, Xinyang Dong, Wenbing Huang, Quansheng Wu, Wanjian Yin, Xi Dai, Shuang Jia, Hangtian Zhu, Jiangang Guo, and Hongjian Zhao for useful discussions.

Author contributions

Zhendong Cao and Lei Wang wrote the code and trained the model. Xiaoshan Luo performed the numerical calculations. Jian Lv and Lei Wang designed the research and supervised the project. All authors contributed to the data analysis, discussion of the results, and preparation of the manuscript.

Data availability

We have released the codes and trained model at https://github.com/deepmodeling/CrystalFormer.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2025.09.035.

References

[1] Woodley SM, Catlow R. Crystal structure prediction from first principles. Nat Mater 2008;7:937–46.
[2] Oganov AR, Pickard CJ, Zhu Q, et al. Structure prediction drives materials discovery. Nat Rev Mater 2019;4:331–48.
[3] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4:268–76.
[4] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. Science 2018;361:360–5.
[5] Merchant A, Batzner S, Schoenholz SS, et al. Scaling deep learning for materials discovery. Nature 2023;624:80–5.
[6] Chen C, Nguyen DT, Lee SJ, et al. Accelerating computational materials discovery with artificial intelligence and cloud high-performance computing: from large-scale screening to experimental validation. 2024; arXiv: 2401.04070.
[7] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023; arXiv: 2303.08774.
[8] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. ICML 2021;139:8821–31.
[9] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). p. 10684–95.
[10] Ingraham JB, Baranov M, Costello Z, et al. Illuminating protein space with a programmable generative model. Nature 2023;623:1070–8.
[11] Xie T, Fu X, Ganea OE, et al. Crystal diffusion variational autoencoder for periodic material generation. 2021; arXiv: 2110.06197.
[12] Luo Y, Liu C, Ji S. Towards symmetry-aware generation of periodic materials. 2023; arXiv: 2307.02707.
[13] Jiao R, Huang W, Lin P, et al. Crystal structure prediction by joint equivariant diffusion. 2023; arXiv: 2309.04475.
[14] Zheng S, He J, Liu C, et al. Towards predicting equilibrium distributions for molecular systems with deep learning. 2023; arXiv: 2306.0544.
[15] Yang M, Cho K, Merchant A, et al. Scalable diffusion for materials discovery. 2023; arXiv: 2311.09235.
[16] Zeni C, Pinsler R, Zügner D, et al. Mattergen: a generative model for inorganic materials design. 2023; arXiv: 2312.03687.
[17] Xiao H, Li R, Shi X, et al. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. Nat Commun 2023;14:7027.
[18] Flam-Shepherd D, Aspuru-Guzik A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files. 2023; arXiv: 2305.05708.
[19] Antunes LM, Butler KT, Grau-Crespo R. Crystal structure generation with autoregressive large language modeling. 2023; arXiv: 2307.04340.
[20] Gruver N, Sriram A, Madotto A, et al. Fine-tuned language models generate stable inorganic materials as text. 2024; arXiv: 2402.04379.
[21] Luo X, Wang Z, Gao P, et al. Deep learning generative model for crystal structure prediction. 2024; arXiv: 2403.10846.
[22] Ye CY, Weng HM, Wu QS. Con-CDVAE: a method for the conditional generation of crystal structures. 2024; arXiv: 2403.12478.
[23] Zhu R, Nong W, Yamazaki S, et al. WyCryst: wyckoff inorganic crystal generator framework. Matter 2024;7:3469–88.
[24] AI4Science M, Hernandez-Garcia A, Duval A, et al. Crystal-GFN: sampling crystals with desirable properties and constraints. 2023; arXiv: 2310.04925.
[25] Nguyen TM, Tawfik SA, Tran T, et al. Hierarchical gflownet for crystal structure generation. In: Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
[26] Jiao R, Huang W, Liu Y, et al. Space group constrained crystal generation. 2024; arXiv: 2402.03992.
[27] Glazer M, Burns G, Glazer A. Space Groups for Solid State Scientists. Oxford: Elsevier; 2012.
[28] Urusov VS, Nadezhina TN. Frequency distribution and selection of space groups in inorganic crystal chemistry. J Struct Chem 2009;50:22–37.
[29] Marsh RE. P1 or $P\bar{1}$? or something else? Acta Crystallogr-Sect B 1999;55:931–6.
[30] Cheetham AK, Seshadri R. Artificial intelligence driving materials discovery? Perspective on the article: scaling deep learning for materials discovery. Chem Mater 2024;36:3490–5.
[31] Hornfeck W. On the combinatorics of crystal structures: number of wyckoff sequences of given length. Acta Crystallogr Sect A: Found Adv 2022;78:149–54.
[32] Oganov AR, Glass CW. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. The J Chem Phys 2006;124:244704.
[33] Davies DW, Butler KT, Jackson AJ, et al. Computational screening of all stoichiometric inorganic materials. Chem 2016;1:617–27.
[34] Zhao Y, Cui Y, Xiong Z, et al. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. ACS Omega 2020;5:3596–606.
[35] Liang H, Stanev V, Kusne AG, et al. Cryspnet: crystal structure predictions via neural networks. Phys Rev Mater 2020;4:123802.
[36] Wang DY, Lv HF, Wu XJ. Crystallographic groups prediction from chemical composition via deep learning. Chin J Chem Phys 2023;36:66–74.
[37] Venkatraman V, Carvalho PA. Accurate space-group prediction from composition. J Appl Crystallogr 2024;57:975–85.
[38] Hahn T, Shmueli U, Arthur JW. International Tables for Crystallography, Vol. 1. Dordrecht: Reidel Dordrecht; 1983.
[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems.
[40] Wirnsberger P, Ballard AJ, Papamakarios G, et al. Targeted free energy estimation via learned mappings. J Chem Phys 2020;153:144112.
[41] Xie H, Zhang L, Wang L. m^* of two-dimensional electron gas: a neural canonical transformation study. SciPost Phys 2023;14:154.
[42] Parthé E, Gelato LM. The standardization of inorganic crystal-structure data. Acta Crystallogr Sect A 1984;40:169–83.
[43] Hautier G, Fischer C, Ehrlacher V, et al. Data mined ionic substitutions for the discovery of new compounds. Inorg Chem 2011;50:656–63.
[44] Glawe H, Sanna A, Gross EKU, et al. The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining. New J Phys 2016;18:093011.
[45] Wang HC, Botti S, Marques MAL. Predicting stable crystalline compounds using chemical similarity. npj Comput Mater 2021;7:12.
[46] Jain A, Bligaard T. Atomic-position independent descriptor for machine learning of material properties. Phys Rev B 2018;98:214112.
[47] Zhou Q, Tang P, Liu S, et al. Learning atoms for materials discovery. Proc Natl Acad Sci USA 2018;115:E6411–7.
[48] Chen C, Ye W, Zuo Y, et al. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater 2019;31:3564–72.
[49] Zhang D, Bi H, Dai FZ, et al. Dpa-1: pretraining of attention-based deep potential model for molecular simulation. 2022; arXiv: 2208.08236.
[50] Wang AYT, Mahmoud MS, Czasny M, et al. Crabnet for explainable deep learning in materials science: bridging the gap between academia and industry. Integr Mater Manuf Innov 2022;11:41–56.
[51] Gazzarrini E, Cersonsky RK, Bercx M, et al. The rule of four: anomalous distributions in the stoichiometries of inorganic compounds. npj Comput Mater 2024;10:73.
[52] Palgrave R. An explanation for the rule of four in inorganic materials. ChemRxiv. 2024. https://doi.org/10.26434/chemrxiv-2024-sxqwh.
[53] Alvarez S. Polyhedra in (inorganic) chemistry. Dalton Trans 2005;13:2209–33.

[54] Regnault N, Xu Y, Li MR, et al. Catalogue of flat-band stoichiometric materials. Nature 2022;603:824–8.
[55] Pauling L. The principles determining the structure of complex ionic crystals. J Am Chem Soc 1929;51:1010–26.
[56] Goldschmidt V. Crystal structure and chemical constitution. Trans Faraday Soc 1929;25:253–83.
[57] Pettifor D. Structure maps for. pseudobinary and ternary phases. Mater. Sci Technol 1988;4:675–91.
[58] Fischer CC, Tibbetts KJ, Morgan D, et al. Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 2006;5:641–6.
[59] Allahyari Z, Oganov AR. Coevolutionary search for optimal materials in the space of all possible compounds. npj Comput Mater 2020;6:55.
[60] Maddox J. Crystals from first principles. Nature 1988;335. 201–201.
[61] Wales DJ. Symmetry, near-symmetry and energetics. Chem Phys Lett 1998;285:330–6.
[62] Avery P, Zurek E. Randspg: an open-source program for generating atomistic crystal structures with specific spacegroups. Comput Phys Commun 2017;213:208–16.
[63] Fredericks S, Parrish K, Sayre D, et al. Pyxtal: a python library for crystal structure generation and symmetry analysis. Comput Phys Commun 2021;261:107810.
[64] Wang Y, Lv J, Zhu L, et al. Crystal structure prediction via particle-swarm optimization. Phys Rev B 2010;82:094116.
[65] Pickard CJ, Needs RJ. Ab initio random structure searching. J Phys: Condens Matter 2011;23:053201.
[66] Lyakhov AO, Oganov AR, Stokes HT, et al. New developments in evolutionary structure prediction algorithm uspex. Comput Phys Commun 2013;184:1172–82.
[67] Falls Z, Avery P, Wang X, et al. The xtalopt evolutionary algorithm for crystal structure prediction. J Phys Chem C 2020;125:1601–20.
[68] Cheng G, Gong XG, Yin WJ. Crystal structure prediction by combining graph network and optimization algorithm. Nat Commun 2022;13:1492.
[69] Wang HC, Schmidt J, Marques MAL, et al. Symmetry-based computational search for novel binary and ternary 2d materials. 2D Mater 2023;10:035007.
[70] Zhang Q, Choudhury A, Chernatynskiy A. A symmetry-oriented crystal structure prediction method for crystals with rigid bodies. 2024; arXiv: 2407.21337.
[71] Sun W, Dacek ST, Ong SP, et al. The thermodynamic scale of inorganic crystalline metastability. Sci Adv 2016;2:e1600225.
[72] Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci 2022;2:718–28.
[73] Zhang D, Liu X, Zhang X, et al. DPA-2: towards a universal large atomic model for molecular and material simulation. 2023; arXiv: 2312.15492.
[74] Batatia I, Benner P, Chiang Y, et al. A foundation model for atomistic materials chemistry. 2024; arXiv: 2401.00096.
[75] Vasala S, Karppinen M. A2B'BO6 perovskites: a review. Prog Solid State Chem 2015;43:1–36.
[76] Wang Y, Baldassarri B, Shen J, et al. Landscape of thermodynamic stabilities of A2BB'O6 compounds. Chem Mater 2024;36:6816–30.
[77] Miao N, Zhou H, Mou L, et al. CGMH: constrained sentence generation by metropolis-hastings sampling. AAAI 2019;33:6834–42.
[78] Ong SP, Richards WD, Jain A, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 2013;68:314–9.
[79] Chu IH, Roychowdhury S, Han D, et al. Predicting the volumes of crystals. Comput Mater Sci 2018;146:184–92.
[80] Okabe R, Cheng M, Chotrattanapituk A, et al. Structural constraint integration in generative model for discovery of quantum material candidates. 2024; arXiv: 2407.04557.
[81] Verkuil R, Kabeli O, Du Y, et al. Language models generalize beyond natural proteins. BioRxiv 2022:2022-12.
[82] Meredig B, Wolverton C. A hybrid computational-experimental approach for automated crystal structure solution. Nat Mater 2013;12:123–7.
[83] Parackal AS, Goodall REA, Faber FA, et al. Identifying crystal structures beyond known prototypes from X-ray powder diffraction spectra. 2023; arXiv: 2309.16454.
[84] Chen C, Zuo Y, Ye W, et al. Learning properties of ordered and disordered materials from multi-fidelity data. Nat Comput Sci 2021;1:46–53.
[85] Franceschetti A, Zunger A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. Nature 1999;402:60–3.
[86] Cheng G, Gong XG, Yin WJ. Global optimization in the discrete and variable-dimension conformational space: the case of crystal with the strongest atomic cohesion. 2023; arXiv: 2302.13537.
[87] Zhao XG, Yang JH, Fu Y, et al. Design of lead-free inorganic halide perovskites for solar cells via cation-transmutation. J Am Chem Soc 2017;139:2630–8.
[88] Ren Z, Tian SIP, Noh J, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. Matter 2022;5:314–35.
[89] Zhao Y, Al-Fahdi M, Hu M, et al. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. Adv Sci 2021;8:2100566.
[90] Ahmad R, Cai W. Free energy calculation of crystalline solids using normalizing flows. Modell Simul Mater Sci Eng 2022;30:065007.
[91] Wirnsberger P, Papamakarios G, Ibarz B, et al. Normalizing flows for atomic solids. Mach Learn: Sci Technol 2022;3:025009.
[92] Köhler J, Invernizzi M, De Haan P, et al. Rigid body flows for sampling molecular crystal structures. 2023; arXiv: 2301.11355.
[93] Miller BK, Chen RT, Sriram A, et al. FlowMM: generating materials with riemannian flow matching. 2024; arXiv: 2406.04713.
[94] Gebauer NW, Gastegger M, Schütt KT. Generating equilibrium molecules with deep neural networks. 2018; arXiv: 1810.11347.
[95] Gebauer NWA, Gastegger M, Schütt KT. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, 2019.
[96] Gebauer NWA, Gastegger M, Hessmann SSP, et al. Inverse design of 3d molecular structures with conditional generative neural networks. Nat Commun 2022;13:973.
[97] Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. 2023; arXiv: 2305.10601.
[98] Goodall REA, Parackal AS, Faber FA, et al. Rapid discovery of stable materials by coordinate-free coarse graining. Sci Adv 2022;8:eabn4117.
[99] Yan K, Liu Y, Lin Y, et al. Periodic graph transformers for crystal material property prediction. NeurIPS 2022;35:15066–80.
[100] Taniai T, Igarashi R, Suzuki Y, et al. Crystalformer: infinitely connected attention for periodic structure encoding. 2024; arXiv: 2403.11686.
[101] Xu H, Qian D, Wang J. Predicting many properties of crystals by a single deep learning model. 2024; arXiv: 2405.18944.
[102] Rubungo AN, Arnold C, Rand BP, et al. Llm-prop: predicting physical and electronic properties of crystalline solids from their text descriptions. 2023; arXiv: 2310.14029.
[103] Schmidt J, Cerqueira TF, Romero AH, et al. Improving machine-learning models in materials science through large datasets. Mater Today Phys 2024;48:101560.
[104] Cao Z, Wang L. Crystalformer-rl: reinforcement fine-tuning for materials design. 2025; arXiv: 2504.02367.
[105] Groom CR, Bruno IJ, Lightfoot MP, et al. The cambridge structural database. Struct Sci 2016;72:171–9.
[106] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. NeurIPS 2020;33:1877–901.
[107] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. 2020; arXiv: 2001.08361.
[108] Frey NC, Soklaski R, Axelrod S, et al. Neural scaling of deep chemical models. Nat Mach Intell 2023;5:1297–305.
[109] Müller U. Symmetry Relationships between Crystal Structures: Applications of Crystallographic Group Theory in Crystal Chemistry. Oxford: Oxford University Press; 2013.
[110] Stokes HT, Hatch DM. Isotropy Subgroups of the 230 Crystallographic Space Groups. Singapore: World Scientific; 1989.
[111] Cox S, White AD. Symmetric molecular dynamics. J Chem Theory Comput 2022;18:4077–81.
[112] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: a simple approach to controlled text generation. 2019; arXiv: 1912.02164.
[113] Zhang M, Jiang N, Li L, et al. Language generation via combinatorial constraint satisfaction: a tree search enhanced monte-carlo approach. 2020; arXiv: 2011.12334.
[114] Qin L, Welleck S, Khashabi D, et al. Cold decoding: energy-based constrained text generation with langevin dynamics. NeurIPS 2022;35:9538–51.
[115] Lew AK, Tan Z-X, Grand G, et al. Sequential monte carlo steering of large language models using probabilistic programs. 2023; arXiv: 2306.03081.
[116] Lai Q, Yao L, Gao Z, et al. End-to-end crystal structure prediction from powder X-ray diffraction. 2024; arXiv: 2401.03862.
[117] Raymond ES. The Cathedral and the Bazaar: Musings On Linux and Open Source by An Accidental Revolutionary. Sebastopol: O'Reilly Media; 1999.