Physics Guided Generative Adversarial Networks for Generations of Crystal Materials with Symmetry Constraints

Yong Zhao\textsuperscript{1} Edirisuriya M. Dilanga Siriwardane\textsuperscript{1,2} Zhenyao Wu\textsuperscript{1} Ming Hu\textsuperscript{3} Nihang Fu\textsuperscript{1} Jianjun Hu\textsuperscript{3}

\textsuperscript{1}Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201
\{yongz, zhenyao, nihang\}@email.sc.edu

\textsuperscript{2}Department of Physics
University of Colombo
Colombo 3, Sri Lanka
dilanga@phys.cmb.ac.lk

\textsuperscript{3}Department of Mechanical Engineering
University of South Carolina
Columbia, SC 29201
hu@sc.edu

\textsuperscript{4}Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201
\{jianjunh\}@cse.sc.edu

Abstract

Discovering new materials is a long-standing challenging task that is critical to the progress of human society. Conventional approaches such as trial-and-error experiments and computational simulations are labor-intensive or costly with their success heavily depending on experts’ heuristics. Recently deep generative models have been successfully proposed for materials generation by learning implicit knowledge from known materials datasets, with performance however limited by their confinement to a special material family or failing to incorporate physical rules into the model training process. Here we propose a Physics Guided Crystal Generative Model (PGCGM) for new materials generation, which captures and exploits the pairwise atomic distance constraints among neighbor atoms and symmetric geometric constraints. By augmenting the base atom sites of materials, our model can generate new materials of 20 space groups. With atom clustering and merging on generated crystal structures, our method increases the generator’s validity by 8 times compared to one of the baselines and by 143% compared to the previous CubicGAN along with its superiority in properties distribution and diversity. We further validated our generated candidates by Density Functional Theory (DFT) calculation, which successfully optimized/relaxed 1869 materials out of 2000, of which 39.6% are with negative formation energy, indicating their stability.

Keywords: Physics, Symmetry, GAN

1 Introduction

Solid crystal materials such as ion conductors, perovskites, photovoltaics and piezoelectrics, play an important role in modern industries. Over centuries, humanity has dedicated a lot of efforts to discover high-performance functional materials. However, for now, only around 250,000 experimentally determined inorganic materials have been found as collected in the ICSD database [3], which covers a fairly small material space compared to the almost infinite chemical design space considering the number of elements in periodic table and possible symmetries of crystal structures. Traditional trial-and-error tinkering methods for materials discovery are mainly reliant on domain experts’ knowledge [24], which is time- and labor-intensive. To meet the high demand for
new functional materials, we need more efficient strategies to explore the vast chemical space to accelerate the materials discovery process.

One of the most popular approaches to generate new materials is based on element substitution combined with high-throughput virtual screening (HTVS) [28]. The whole process contains three steps: (1) combinatorially substituting elements in known crystal structures, (2) relaxing the candidate structures using density functional theory (DFT) calculations, and (3) experimental verification. Several large computational materials databases created by HTVS are Materials Project (MP) [17] and Open Quantum Materials Database (OQMD) [20]. Despite its promising usage in material design, one fundamental drawback of HTVS is that it cannot generate materials beyond the structural prototypes of existing materials and it heavily depends on experts’ intuitions.

One way to overcome the drawbacks of HTVS to find optimal materials is to perform global optimization using genetic or evolutionary algorithms. Global optimization is used to search an optimal phase of materials by setting an objective function and, thus it can be used to design new materials. Simulated annealing is one of early global optimization algorithms to design atomic configuration of alloys [10] and predict the crystal structure of boron nitride [7]. The minima hopping is proposed to find unknown crystalline structures in several cases by adapting softening process [1] and later on is adjusted to predict magnetic systems [8]. USPEX [11] and CALYPSO [33] are two typical approaches that use evolutionary algorithms. Both find completely new crystal structures beyond materials in existing databases by optimizing the property of stability.

Another promising approach to design solid materials beyond known crystal structure prototypes is generative deep learning models [36], which can learn data distribution (knowledge of forming stable crystal structures) from known materials and then sample from it to generate novel materials. Variational Auto-encoder (VAE) [19] and Generative Adversarial Network (GAN) [12] are two popular generative models used to generate materials. VAE [19] composes of two deep neural networks, an encoder and a decoder. The encoder is trained to encode the materials into latent vectors and the decoder reconstructs the materials from the latent vectors, then use different strategies to generate/reconstruct materials by sampling latent vectors. iMatGen [24] is believed to be the first work that uses VAE to realize the inverse design of solid materials. iMatGen encodes unit cells into 3D grid based representations, and spherical linear interpolation and Gaussian random sampling are used to sample from the latent space to generate new materials. Hoffmann et al. [16] extend iMatGen by combining a UNet part to segment reconstructed 3D voxel images into atoms and assign atoms with a number. Base on iMatGen and Hoffmann et al., ICSG3D [5] integrates formation energy per atom into 3D voxelized solid crystals and enables the VAE to encode materials and energy simultaneously, which makes it possible to generate materials subject to user-defined formation condition. Another approach to represent 3D crystals is to encode 2D crystallographic representations as the combination of real space and reciprocal-space Fourier-transformed features [29]. In CDVAE [35], authors propose to generate materials in a diffusion process [15] in the decoder. The diffusion process moves atoms into positions in the lower energy space to generate stable crystals.

GANs [12] also consist of two deep neural networks, a generator and a discriminator (critic). The generator generates fake materials with inputs of random vectors sometimes conditioning on elements and space groups while the discriminator tries to tell real materials from generated ones. With learnt knowledge of forming crystals, the generator can directly create new materials. The first method to generate materials using GAN is CrystalGAN [25], which leverages a CycleGAN [37] to generate new ternary materials from existing binaries. However, it remains uncertain whether CrystalGAN can be extended to produce more complex crystals. Both GANCSP [18] and CubicGAN [36] use a “point cloud” (containing fractional coordinates, element properties, and lattice parameters) as inputs to build a model that generates crystals conditioned on composition or both composition and space group. The difference between them is GANCSP can only generate structures of the Mg-Mn-O system but CubicGAN can generate more diverse systems under three space groups. In CCDCGAN [21], the authors use 3D voxelized crystals as inputs for an autoencoder. The autoencoder converts them to 2D crystal graphs, which is used as the inputs to the GAN model. A formation energy based constraint module is trained with the discriminator. This module automatically guides the searching for local minima in the latent space. Less related works include MatGAN [6] and CondGAN($x^{bp}$) [30] developed for generating only chemical formulas.

Despite the success of VAEs and GANs in material generation [24, 35, 36], current generative models all have several major drawbacks. For example, the iMatGen algorithm [24] can only focus on a specific chemical system such as vanadium oxides and only several new metastable VxOy materials were discovered out of 20,000 generated hypothetical materials. Similarly, in [18] and [25], their models mainly generate Mg-Mn-O ternary materials and hydride systems, respectively. VAE-UUnet pipeline developed in [5] expands the diversity of generated materials and can reconstruct the atom coordinates more accurately by incorporating UNet segmentation and conditioning on properties. However, VAE-UUnet still confines itself to cubic crystal system generation and the number of atoms in a unit cell is restricted no more than 40. All above discussed works do not realize high-throughput generation of materials. CubicGAN [36] is the first model that achieves large-scale generation of new (cubic) materials, which has discovered four new prototypes with 506 materials confirmed to be stable by DFT. Although CubicGAN
has generated millions of crystal structures with hundreds of stable ones confirmed, the generated structures are limited to three space groups in the cubic crystal system. Moreover, it is not capable of generating generic atom coordinates. While these works open the door to generative design of new materials, several unique challenges still remain that prevents effective generative design: (1) how to learn the physical atomic constraints of stable materials to enable efficient sampling; (2) how to achieve precise generation of atom fractional coordinates and lattice parameters; (3) how to handle the extreme bias of the distribution of materials in 230 space groups. In this work, we introduce a new physics guided crystal generative model (PGCGM) to exploit the physical rules for addressing aforementioned challenges. Our contributions are summarized as follows:

1. We present a new physics guided deep generative model for crystal generation that combines the space group affine transformation and an efficient self-augmentation method.

2. We propose two physics-oriented losses based on atomic pairwise distance constraints and symmetry to fuse the physical laws into deep learning model training.

3. We evaluate our model against two baselines to show its superiority and perform DFT calculations to validate our generated structures with high success rate (93.5% can be optimized successfully).

2 Results

Problem statement and notations  The structure of an inorganic material can be represented by a unit cell in material science. The unit cell is the smallest unit that completely reflects the arrangement of atoms in the 3D space. Additionally, the unit cell describes the periodic structure of an inorganic material and it can be repeated infinite times along three directions to form a super cell. A material \( \mathcal{M} \) can be denoted as following:

\[
\mathcal{M} = (E, B, P, O),
\]

where: (a) \( E = (e_0, e_1, e_2) \in \mathbb{E} \) denotes elements in materials, where \( \mathbb{E} \) is the element set in periodic table. In this work, we only deal with ternary materials so that there are only three unique elements in the unit cell;

(b) \( B = (b_0, b_1, b_2) \in \mathbb{R}^{3 \times 3} \) denotes the symmetry equivalent positions termed as base atom sites. \( b_i \) is fractional coordinates of an atom denoted by \([u, v, w]^T\). We choose materials that one element only has one base atom site so that three atom sites can be used to represent the atom positions. Moreover, any one atom site of each element can be considered as the base atom site for that element;

(c) \( P = (a, b, c, \alpha, \beta, \gamma) \in \mathbb{R}^6 \) are six lattice parameters that define three lengths and three angles of the unit cell;

(d) \( O = (t_0, t_1, \ldots, t_n) \in \mathbb{R}^{4 \times 4} \) denotes affine matrix that represents the symmetry operations defined by space groups \( sgp \). \( t_j \) is one affine operator containing the rotation and translation matrices. \( n \) is determined by space groups. Generally the higher symmetry of a space group, the larger \( n \). \( n \) can be as small as 1 or as large as 192.

Now we can model the generation of materials as follows:

\[
(B, P) = f_\theta(Z, E, sgp),
\]

where \( f_\theta \) is the generative model that learns the knowledge of forming crystal structures given inputs of random noise \( Z \), element set \( E \), and space group \( sgp \).

Physics guided crystal generative model  We describe Physics Guided Crystal Generative Model (PGCGM) in Figure 1. The PGCGM mainly consists of four major components: (1) discriminator, (2) generator, (3) self-augmentation, and (4) atom distance matrix/loss calculation module. In the generator and discriminator, affine matrix is integrated into the training to generate fake materials and tell fake materials from real ones, respectively. Affine matrix is related to symmetry information for space groups. The implicit combination of affine matrix and base atom sites can help keep the symmetry when generating materials. Self-augmentation increases the training materials for underrepresented space groups by randomly forming base atom sites. With three sets of base atom sites, we can not only have a fixed size of input to the discriminator, but also deduce more physical information for crystals to help the discriminator better distinguish real materials from fake ones. Furthermore, we design two kinds of physics guided losses. Any set of base atom sites can be converted to unique full atom sites when generating three sets of base atom sites, it implicitly can be stated that the three sets of base atom sites should be different but the full atom sites converted from them separately ought to be same. Hence a specific loss is invented to explicitly incorporate this rule into training of the generator. In order to restrict the two atoms in the 3D space
to be not too close or not too distant, inter- and intra-atom distance loss are proposed. With distance loss, the generator further can generate reasonable lattice parameters in order to push any pair of atoms to fall into a certain range.

Figure 1: The overview of physics guided crystal generative model (PGCGM). The PGCGM comprises four components. 
(a) The generator takes affine matrix $O$, random noise $Z$, and element properties $E$ as inputs. The affine matrix and random noise are projected to two vectors by 2D convolutional networks and fully connected layers, respectively and then the two vectors are merged and projected to generate lattice parameters $P^*$ by fully connected layers. The element properties are projected to a vector by 1D convolutional networks and then it is merged with the vector projected from random noise to generate three sets of base atom sites ($B^0_{fake}, B^1_{fake}, B^2_{fake}$). (b) The discriminator has two input branches. It shares with the same affine matrix branch as in the generator. The assembled crystal representation matrix from base atom sites, lattice parameters, and properties calculated from them as the input to 2D convolutional networks. The assembled matrix is zero-padded to form a matrix with shape of $3 \times 8 \times 8$. (c) The self-augmentation performed on the base atom sites. We choose three sets of base atom sites from three elements randomly and with space group, we can calculate more crystal information to assemble the input matrix for the discriminator. (d) Inter- and intra-atom distance matrices are calculated from three sets of base atom sites for both real and fake materials. Then we design distance based losses to constrain the distance between two atoms in a certain range as shown in the grey area form by two circles.

Physics guided loss function The original GAN [12] is notoriously hard to train because of saturation and mode collapse in discriminator. We take advantage of WGAN-GP [13] with gradient penalty to enhance the training stability in our network. WGAN-GP changes the Sigmoid function of the discriminator to a 1-Lipschitz function while introducing a gradients penalty term to enforce the norm of gradients to be close to 1. The loss function is described in Eq. (3):

$$\hat{M}^* = \epsilon M_{fake}^* + (1 - \epsilon) M_{real}^*, \quad \epsilon \sim U(0, 1),$$

$$L_{dis} = D(M_{fake}^*) - D(M_{real}^*) + \lambda_d \left( \| \nabla_\hat{M}^* D(\hat{M}^*) \|_2 - 1 \right)^2,$$

$$L_{adv} = -D(M_{fake}^*),$$

where $\hat{M}^*$ is linearly interpolated between real and fake materials and $\epsilon$ is uniformly sampled from 0 and 1. $L_{dis}$ and $L_{adv}$ represent the loss function of the discriminator and adversarial loss for generator respectively. The third term in $L_{dis}$ is the gradient penalty and $\lambda_d$ is set 10. $D(.)$ means the score result from the discriminator.

Atom Distance Losses To ensure that the atoms in generated crystal structures are not crowded or not too far
apart from each other, we introduce the inter- and intra-atom distance based losses as following:

\[
\mathcal{L}_{\text{inter}} = \frac{1}{N*9} \sum_{i=1}^{N} \left[ \max(H_{\text{inter}}, \phi_{\text{inter}}^\text{upper} S_{\text{inter}}) - \phi_{\text{inter}}^\text{upper} S_{\text{inter}} \right]^2
+ \left[ \min(H_{\text{inter}}, \phi_{\text{inter}}^\text{lower} S_{\text{inter}}) - \phi_{\text{inter}}^\text{lower} S_{\text{inter}} \right]^2,
\]

\[
\mathcal{L}_{\text{intra}} = \frac{1}{N*9} \sum_{i=1}^{N} \left[ \max(H_{\text{intra}}, \phi_{\text{intra}}^\text{upper} S_{\text{intra}}) - \phi_{\text{intra}}^\text{upper} S_{\text{intra}} \right]^2
+ \left[ \min(H_{\text{intra}}, \phi_{\text{intra}}^\text{lower} S_{\text{intra}}) - \phi_{\text{intra}}^\text{lower} S_{\text{intra}} \right]^2,
\]

where \( \mathcal{L}_{\text{inter}} \) constrains the distance in \( H_{\text{inter}} \). \( \max(H_{\text{inter}}, \phi_{\text{inter}}^\text{upper} S_{\text{inter}}) - \phi_{\text{inter}}^\text{upper} S_{\text{inter}} \) enforces the atom distance to be smaller than \( \phi_{\text{inter}}^\text{upper} S_{\text{inter}} \), and \( \min(H_{\text{inter}}, \phi_{\text{inter}}^\text{lower} S_{\text{inter}}) - \phi_{\text{inter}}^\text{lower} S_{\text{inter}} \) enforces the atom distance to be bigger than \( \phi_{\text{inter}}^\text{lower} S_{\text{inter}} \). \( S_{\text{inter}} \) are atom radius sum corresponding to each pair of atoms in \( H_{\text{inter}} \) and \( \phi_{\text{inter}} \) are control weights for upper and lower bound of inter-atom distance, respectively.

In this way, the distance of two atoms is constrained to be in the grey area indicated by two circles in sub-figure (d) of Figure 1. Similarly, \( \mathcal{L}_{\text{intra}} \) constrains the distance in a range in \( H_{\text{intra}} \). \( N \) is batch size and 9 is the number of distance value in \( H_{\text{inter}} \) and \( H_{\text{intra}} \).

Base and Average Full Coordinates Losses The generator generates three sets of base atom sites \( (B_{\text{fake}}^0, B_{\text{fake}}^1, B_{\text{fake}}^2) \) which are supposed to be different but a part of full coordinates. The averaged transformation to \( (F_{\text{fake}}^0, F_{\text{fake}}^1, F_{\text{fake}}^2) \) from base atom sites should be exactly same. With these implicit rules, we design two losses to explicitly enforce them in the generator as expressed below:

\[
\mathcal{L}_{\text{full}} = \frac{1}{N*9} \sum_{i=1}^{N} \left[ \max(0, \cos(F_{\text{fake}}^0 \Phi_{\text{fake}}^0, F_{\text{fake}}^1 \Phi_{\text{fake}}^1, F_{\text{fake}}^2 \Phi_{\text{fake}}^2)) \right]
+ \max(0, \cos(F_{\text{fake}}^0 \Phi_{\text{fake}}^0, F_{\text{fake}}^1 \Phi_{\text{fake}}^1, F_{\text{fake}}^2 \Phi_{\text{fake}}^2)),
\]

\[
\mathcal{L}_{\text{base}} = \frac{1}{N*9} \sum_{i=1}^{N} \left[ (1 - \cos(B_{\text{fake}}^0 \Phi_{\text{fake}}^0, B_{\text{fake}}^1 \Phi_{\text{fake}}^1, B_{\text{fake}}^2 \Phi_{\text{fake}}^2)) \right]
+ (1 - \cos(B_{\text{fake}}^0 \Phi_{\text{fake}}^0, B_{\text{fake}}^1 \Phi_{\text{fake}}^1, B_{\text{fake}}^2 \Phi_{\text{fake}}^2)),
\]

where \( \cos \) is cosine similarity function. We normalize each coordinate value across the mini-batch of size \( N \). 9 is the number of coordinates.

Full Generator Loss By combining above losses, we can achieve our full loss for the generator:

\[
\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{inter}} + \lambda_2 \mathcal{L}_{\text{intra}} + \lambda_3 \mathcal{L}_{\text{full}} + \lambda_4 \mathcal{L}_{\text{base}}
\]
**Dataset** We collect our material data from MP [17], ICSD [3] and OQMD (v1.4) [20]. In total, 42072 ternary materials with 20 space groups are curated when we start this project. We use a 80-20 random training/validation split for all of our experiments. We term the dataset with 42072 materials as **MIO**. When conducting this project, the newest version of OQMD is just yet released. There are 9441 ternary materials that are filtered by the same criteria and are brand new materials in the new OQMD (v1.5). We use these 9441 ternary materials as our test dataset **TST** to compare our method with two baselines. Details regarding dataset collection are in Dataset Curation section of supplementary materials.

**Generation performance** We compare PGCGM with two latest algorithms that can generate crystals with multiple chemical systems instead of only a special group of materials, such as VxOy and Mg-Mn-O systems [18, 24]. FTCP [29] combines real space properties (e.g., atom coordinates) and momentum-space properties to represent crystal structures. Then a CNN based VAE is trained for materials generation. CubicGAN [36] trains a WGAN-GP [13] to generate cubic structures in three space groups and here we expand the original method to 20 space groups.

| Method          | Validity (%) | Prop. Dist. | (%)
|-----------------|--------------|-------------|
|                 | CIFs | Distance | Charge | minD | maxD | density | Diversity |
| FTCI            | 0.88 | 63.28    | 49.89   | 1.685 | 0.754 | 2.895   | 89.9     |
| CubicGAN        | 4.97 | 99.0    | 59.47   | 0.626 | 3.476 | 3.871   | 98.0     |
| PGCGM           | 1.98 | **99.54** | 57.36   | **0.224** | 3.664 | 2.675   | **98.4** |
| PGCGM+dist      | 7.14 | 99.47   | 61.82   | 0.405 | 0.520 | 0.765   | 96.3     |
| PGCGM+dist+coor | 6.07 | 99.43   | **63.34** | 0.357 | **0.490** | 0.791   | 97.0     |

Table 1: Material generation performance.

The performance is shown in Table 1. For each method, we sample 500,000 structures and the percentage of Crystallographic Information Files (CIFs) that are readable by pymatgen [26] are shown in the CIFs column. For PGCGM and CubicGAN, we perform atom clustering and merging. We can find that PGCGM+dist has the largest percentage of materials left and PGCGM+dist+coor comes next. It might tell us that distance and coordinates losses play a big part in generating readable materials. For later percentage related metrics, we use the number of CIFs left of each method as denominator. Our model significantly outperforms FTCP by 36.4% in terms of distance validity and is slightly better than CubicGAN. In terms of distance validity, our model outperforms FTCP and CubicGAN by 6.5% and 27.0%, respectively. Since validity are relatively weak metrics, property distribution is further used to provide a stronger metric to evaluate whether the generated materials are realistic. Our model significantly outperforms both two baselines. In terms of minimum atom distance, PGCGM decreases wasserstein distance (WD) by 1.461 compared to FTCP and by 0.402 compared to CubicGAN. In terms of maximum atom distance, PGCGM+dist+coor decreases WD by 0.264 compared to FTCP and by 2.986 compared to CubicGAN. Although CubicGAN has a close minimum atom distance distribution as PGCGM, the much bigger gap of maximum atom distance distribution between CubicGAN and PGCGM+dist+coor indicates that CubicGAN tends to generate large crystal structures. In terms of density, PGCGM+dist decreases WD by 2.130 compared to FTCP and by 3.106 compared to CubicGAN. PGCGM also achieves the best diversity score even though it generates more readable CIFs than FTCP, which further shows that FTCP is not able to generate not only physically realistic materials but also materials with restricted diversity of formulas. We choose PGCGM+dist+coor as our finalized model to generate materials for late analysis since PGCGM+dist+coor has better properties distribution performance than PGCGM and PGCGM+dist on average.

**Analysis of materials optimized by Bayesian Optimization with Symmetry Relaxation** Bayesian optimization with Symmetry Relaxation (BOWSR) algorithm [38] is an approach that uses Bayesian optimization to iteratively search lower energy surface to optimize the crystal structures based on the properties predicted by deep learning methods, such as CGCNN [34] and MEGNet [4]. Instead of expensive DFT for generated materials, we use BOWSR to optimize materials generated by our method and two baseline models. We randomly select 2,000 generated materials with less than and equal to 32 atoms for FTCP, CubicGAN and PGCGM. We select 100 materials for 20 space groups equally generated by PGCGM. Note that we also use the same 2,000 materials of PGCGM for further DFT analysis. Because some space groups are underrepresented (with less than 100 materials) in CubicGAN-generated materials, we select all materials under these space groups and then we select materials for the rest of space groups proportionally to obtain 2000 materials. For FTCP, all materials that can be successfully analyzed to have space groups by pymatgen [26] surprisingly all belong to space group P1, which means FTCP loses the significant symmetric constraints when generating materials. Our method PGCGM
and CubicGAN are much better than FTCP in terms of space groups retention. Moreover, it takes more than 10 times time to optimize materials generated by FTCP than by PGCGM and CubicGAN using BOWSR. We use StructureMatcher from pymatgen [26] to match the generated materials with the corresponding optimized materials by BOWSR.

| Method      | # of optimized materials | Match rate (%) | RMS   |
|-------------|--------------------------|----------------|-------|
| FTCP        | 1982                     | 0.8            | 0.006 |
| CubicGAN    | 1992                     | 26.0           | 0.035 |
| Our approach| 1994                     | 34.9           | 0.052 |

Table 2 shows the match rate and RMS displacement. The match rate is the percentage of materials satisfying the criteria $\text{ltol}=0.2$, $\text{stol}=0.2$, $\text{angle}_{\text{tol}}=0.5$ and then we calculate RMS displacement for the matched materials. Firstly we find that our method has a slightly higher number of successfully optimized materials by BOWSR. However, our method significantly outperforms FTCP and CubicGAN by 4200% and 34.23% in terms of match rate, respectively. It seems that FTCP has the best RMS displace but the extremely low match rate might tell us that BOWSR is hard to optimize materials with low symmetry, such as space group $P\bar{1}$. CubicGAN comes next in terms of RMS displacement.

**Analysis of rediscovering materials in training and test datasets** It would be helpful to show how fast our model can rediscover materials in training dataset MIO and test dataset TST. To do this, we sample different number of materials and then calculate the percentage of materials rediscovered in generated materials. "Anonymized Formula - Space Group ID - # of Atoms” is defined as prototype to identify unique materials in the existing and generated materials. Figure 2a shows the change of unique crystals and rediscovery rate over size of sampling materials. We start to sample materials from half million and the number ends at 60 million eventually. It is found that the percentage of unique materials (cyan line) are decreasing and gradually tend to grow flat as number of sampling materials increases. The rediscover rate of MIO (orange bars) increases consistently over the sampling process and it soars quickly to 42.6% when 35 million materials are sampled. Starting from 42.6%, the percentage of rediscovered materials in MIO grows smoothly and it hit 52.0% with sampling size of 60 million. Similar growing patterns can be observed for the rediscover rate for the test dataset, as shown by blue line in Figure 2a. The rediscover rate reaches 43% at the end of 60 million sampling size. This percentage is lower than that of training dataset because of the different proportions over the 20 space groups in training and test datasets as shown in Supplementary Table S1 and S2.

![Figure 2](image)

(a) The rediscover rate in training and test datasets.

(b) The parity plot for lattice lengths of generated and matched materials.

**Figure 2**: Analysis of rediscovered materials. (a) The discovery rate of prototypes in MIO and TST. The number of unique prototypes is 26,463 (42,072) in MIO and the number of unique prototypes is 8,103 (9,441) in TST. (b) The parity plot for lattice lengths between generated materials and materials that match generated materials in MIO and TST. Top row is for the training dataset and bottom row is for the test dataset (OQMD v1.5), respectively. $R^2$ and $RMSE$ are also used to evaluate the performance of generated lattice lengths compared to existing ones.

After rediscovering the materials in MIO and TST from the generated materials when sampling size is 60 million, we utilize StructureMatcher from pymatgen [26] to test whether the generated materials match the rediscovered materials and to calculate RMS displacement between two matched structures considering all invariances of materials. Because one prototype might correspond to multiple structures in existing and generated
materials, we only show the least RMS displacement by exhausting each pair of existing and generated materials for this prototype. The match rate is the percentage of materials satisfying the criteria $l_{tol}=0.2$, $s_{tol}=0.3$, $angle_{tol}=5.0$. The match rate and RMS are 25.4% and 0.05 for training dataset and are 17.7% and 0.085 for test dataset, respectively. Figure 2b shows parity plot that compares generated lattice lengths against DFT calculated lattice lengths. Surprisingly, the co-relation between the discovered materials in test dataset and generated materials is better than in training dataset in terms of $R^2$. The $R^2$ for lattice a, b, and c in test dataset are 0.606, 0.616, and 0.606, respectively as in Figure 2b, which increases $R^2$ as in training dataset by a factor of six except for lattice c. The rediscovered materials in training dataset have larger lattice a and b and we find that these materials mostly are with cubic space groups. It seems that our approach tends to generate more realistic lattice for non-cubic space groups than cubic space groups.

Figure 3: The distribution of formation energy for 1863 materials and energy above hull for 1579 materials. (a) 39.6% materials are with negative formation energy. (b) Three materials are with energy above hull equal to zero and 106 ones with energy above hull less than 0.25 eV/atom among 1579 materials.

DFT verification We use the same 2,000 materials as in Bayesian optimization for DFT verification. Out of 2000 generated crystals, 93.5% (1869) are successfully optimized, which is significantly better than 33.8% of CubicGAN as reported in [36]. Figure 3a demonstrates the distribution of formation energy of successfully optimized materials after removing 6 materials with formation energy larger than 10 eV. It is observed that most of the materials have formation energy around 0 eV and 39.6% of them have negative formation energy. Negative formation energy indicates potentially stable materials. Figure 3b shows the distribution of 1,579 materials that have energy above hull after removing one material with super large energy above hull. Energy above hull is a stronger indicator whether the materials are stable or not. Overall, 3 materials with energy above hull of 0 eV/atom and 106 (5.3%) ones with energy above hull less than 0.25 eV/atom, which further indicates our model can generate reliably stable materials. All the optimized materials are included in the supplementary materials.

Pair-wise atom distance based loss not only constrains the two atom in a reasonable range, but also helps generate lattice lengths close to DFT-calculated ones. To demonstrate this, we calculate relative error, $R^2$, $RMSE$, and $O$ (outliers percentage) for lattice lengths for 1,869 materials as shown in left panel of Figure 4 and for only 293 cubic materials by PGCGM and 14,432 cubic materials by CubicGAN as shown in right panel of Figure 4. In terms of relative error, we can find that the mean relative error of lattice lengths is much more close to zero regardless of when comparing 1,869 materials or just cubic materials by PGCGM with cubic materials by CubicGAN, which indicates that PGCGM tends to generated precise lattice lengths. In addition, the outliers of lattice lengths in 1,869 materials by PGCGM scatter across 100% and cubic materials from 1,869 ones only have two outliers compared to CubicGAN whose outliers cluster near to 150% even though CubicGAN overall has a lower outliers percentage. We also evaluate the lattice lengths generation performance between PGCGM and CubicGAN with $R^2$ and $RMSE$. For 1,869 materials, the generated lattice lengths by PGCGM scatter across 100% and $RMSE$. For 1,869 materials, the generated lattice lengths by PGCGM better fit to the DFT calculated lattice lengths than CubicGAN in terms of $R^2$. In terms of $RMSE$, PGCGM is generally slight better CubicGAN. When only comparing cubic materials, PGCGM significantly outperforms CubicGAN in terms of both $R^2$ and $RMSE$. Although it is not a direct comparison between PGCGM and CubicGAN, all this findings indicate that our model can generate high quality materials with reasonable lattice lengths.
Figure 4: The comparison of lattice parameters $P^*$ generation performance for 20 space groups and space groups in cubic system. Lattice angles are constrained to fixed values by virtue of crystal systems of 20 space groups. Relative error is calculate by $(\text{length}_{GEN} - \text{length}_{DFT})/\text{length}_{DFT}$, where $\text{length}_{GEN}$ is the generated lattice length and $\text{length}_{DFT}$ is the relaxed lattice length. The bounding boxes correspond to each box plot above them and $R^2$, $RMSE$, and $O$ are used to evaluate the lattice lengths generation performance. $O$ means the percentage of outliers in the box plots. (a) The error distribution of three lattice lengths for 1869 materials generated/relaxed in PGCGM. (b) The error distribution of one lattice length for cubic materials generated/relaxed in PGCGM and CubicGAN, respectively. There are 293 cubic materials optimized in PGCGM and 14,432 cubic materials optimized in CubicGAN successfully.

In Table 3, 20 structures with lowest formation energy are selected for 20 space groups as in the dataset. Before any post-processing, both materials has a large number of atoms. It is easily found that the atoms of the same elements are crowded together in column of $GEN$. After clustering and merging the atoms of the same elements, the number of atoms drop rapidly, such as space group 164 (from 36 to 14) and space group 227 (from 432 to 32) as shown in column of $MER$. Column $OPT$ shows the crystal structures after DFT optimization. Columns of Formula ($GEN$) and Fromula ($MER$) show the formulas before and after clustering and merging.
Table 3: 20 example optimized crystals with lowest energy for 20 space group. GEN: generated; #: no. of atoms; MER: merged; OPT: optimized; FE: formation energy; SG: space group.

| Formula (GEN) | GEN | # | Formula (MER) | MER | OPT | # | FE(eV) | SG |
|---------------|-----|---|---------------|-----|-----|---|--------|----|
| RbScN         | 36  |   | RbScN12       |     |     | 14| -4.400 | 164|
| Cu4BN4        | 432 |   | CuBN2         |     |     | 32| -1.711 | 227|
| Al(BrN)2      | 36  |   | Al(BrN4)2     |     |     | 22| -3.659 | 139|
| HfNF          | 36  |   | HfNF3         |     |     | 16| -3.658 | 186|
| Ti2VN2        | 32  |   | TiVN8         |     |     | 20| -3.503 | 129|
| HfNCl         | 60  |   | HfN2Cl3       |     |     | 30| -3.142 | 194|
| HfBi2F2       | 40  |   | HfBiF8        |     |     | 10| -3.314 | 123|
| Ti2GeN2       | 120 |   | Ti(GeN8)3     |     |     | 28| -2.579 | 221|
| ScInN2        | 32  |   | Sc2InN8       |     |     | 22| -3.690 | 71 |
| Zr2VN2        | 60  |   | Zr(VN4)3      |     |     | 16| -3.666 | 191|
| TiNF          | 24  |   | Ti(NF)2       |     |     | 20| -3.819 | 62 |
| Mg4NF4        | 216 |   | MgNF6         |     |     | 32| -2.267 | 216|
| ZrMnO         | 108 |   | ZrMn2O6       |     |     | 27| -3.487 | 166|
| BeNO          | 48  |   | BeNO          |     |     | 24| -3.721 | 63 |
| Sc(HO2)2      | 56  |   | Sc2HO4        |     |     | 28| -3.672 | 141|
| CaCdF         | 48  |   | Ca2CdF4       |     |     | 28| -3.494 | 122|
| Hf4Os8F       | 312 |   | HfOsF6        |     |     | 32| -3.178 | 225|
| Mn2Sn2N       | 80  |   | MnSnN4        |     |     | 28| -2.611 | 140|
| ZrPoN         | 54  |   | ZrPoN6        |     |     | 24| -3.945 | 148|
| MgZrO         | 108 |   | MgZrO3        |     |     | 30| -4.340 | 167|

**Analysis of three example stable materials** We discovered three compounds with Mg$_2$GaIr, SrYO$_6$ and ZnTe$_2$S$_6$ chemical formulas, which are thermodynamically stable with negative formation energies and zero energy above hull. The structures we found have P6$_3$/mmc [194] (hexagonal), Pm-3 [200] (cubic) and [148] (trigonal) space group symmetries [space group number] (crystal systems) for Mg$_2$GaIr, SrYO$_6$ and ZnTe$_2$S$_6$ compounds, respectively. Figure 5 (a) shows the structures of those three materials. The lattice parameters for Mg$_2$GaIr material are
\( a = b = 4.38 \text{ Å}, c = 8.54 \text{ Å}, \alpha = \beta = 90^0 \) and \( \gamma = 120^0 \), while that for SrYO\(_6\) material are \( a = b = c = 4.61 \text{ Å} \) and \( \alpha = \beta = \gamma = 90^0 \). Moreover, the lattice parameters for ZnTe\(_2\)S\(_6\) material are \( a = b = c = 7.57 \text{ Å} \) and \( \alpha = \beta = \gamma = 49.38^0 \). As shown in Table 4, our spin-polarized DFT calculations show that the both Mg\(_2\)GaIr and ZnTe\(_2\)S\(_6\) compounds have the non-magnetic ground states, whereas SrYO\(_6\) material has a ferromagnetic ground state with a total magnetic moment of 1 \( \mu_B \). Figure 5 (b) contains the electronic band structures for each stable material. It is clear that both Mg\(_2\)GaIr and ZnTe\(_2\)S\(_6\) compounds are metals. However, we can see spin-splitting in SrYO\(_6\) ferromagnetic material. In this compound only spin-down electrons cross the Fermi level, while spin-up electrons have a band gap of 3.09 eV. Thus, this is a half-metal where spin-down electrons show metallic character, while spin-up electrons are insulating. Half metallicity is widely investigated for spintronics and it is vital for developing memory devices and computer processors [27].

Table 4: The electronic and magnetic properties of the stable materials. The material type (metal or semiconductor), band gap, magnetic ground state: GS (non magnetic: NM or ferromagnetic: FM) and the magnetic moment: \( \mu \) are reported for each stable material. For the half-metal, the band gaps for both spin types (Up and Down) are mentioned.

| Material     | Material Type | Band Gap (eV) | Magnetic GS | \( \mu \) (\( \mu_B \)) |
|--------------|---------------|--------------|-------------|--------------------------|
| Mg\(_2\)GaIr | Metal         | 0            | NM          | 0                        |
| SrYO\(_6\)   | Half-metal    | Up:3.09, Down:0.00 | FM | 1                        |
| ZnTe\(_2\)S\(_6\) | Metal       | 0            | NM          | 0                        |

Table 5: The elastic constants \( (C_{ij}) \), bulk (\( B \)) modulus, shear (\( G \)) modulus and Young’s (\( Y \)) modulus in GPa and Poisson’s ratio (\( \nu \)) for the stable materials.

| Material     | \( C_{11} \) | \( C_{12} \) | \( C_{13} \) | \( C_{44} \) | \( C_{66} \) | \( B \) | \( G \) | \( Y \) | \( \nu \) |
|--------------|--------------|--------------|--------------|--------------|--------------|--------|--------|--------|--------|
| Mg\(_2\)GaIr | 177.32       | 98.68        | 132.77       | 52.04        | 36.276       | 36.276 | 96.44 | 40.67  | 0.315  |
| SrYO\(_6\)   | 181.61       | 60.18        | 181.61       | 60.18        | 37.951       | 37.951 | 100.66| 45.85  | 0.302  |
| ZnTe\(_2\)S\(_6\) | 83.13 | 31.49 | 4.56 | 0.85 | 0.382 | 25.821 | 15.34 | 7.69 | 0.285 |

Table 5 contains the elastic constants and the mechanical properties of the stable materials. The elastic stability criteria (Born criteria) for Mg\(_2\)GaIr with P6\(_3\)/mm space group symmetry are \( C_{11} > |C_{12}|, 2 * C_{13}^2 < C_{33}(C_{11} + C_{12}) \), and \( C_{44} > 0 \). The Born criteria for SrYO\(_6\) with Pm-3 space group symmetry are \( C_{11} - C_{12} > 0, C_{11} + 2C_{12} > 0, C_{44} > 0 \), and that for ZnTe\(_2\)S\(_6\) with R-3 space group symmetry are \( C_{11} > |C_{12}|, C_{13}^2 < 0.5 * C_{33}(C_{11} + C_{12}) \) and \( C_{14}^2 + C_{15}^2 < 0.5 * C_{44} * (C_{11} - C_{12}) \), and \( C_{44} > 0 \) [22]. It is clear that all those three materials comply with their elastic stability criteria implying they are mechanically stable. Table 5 also has the bulk (\( B \)) modulus, shear (\( G \)) modulus and Young’s (\( Y \)) modulus and Poisson’s ratio (\( \nu \)) for Mg\(_2\)GaIr, SrYO\(_6\), and ZnTe\(_2\)S\(_6\) compounds. We used Hill approximation as implemented in vaspkit code [14, 32]. It is clear that Mg\(_2\)GaIr and SrYO\(_6\) materials have approximately same \( B \), \( G \), and \( Y \) values, while those values for ZnTe\(_2\)S\(_6\) compound is considerably lower. In Table 5, \( \nu \) of Mg\(_2\)GaIr has the highest value, while lowest \( \nu \) can be obtained from ZnTe\(_2\)S\(_6\). Furthermore, we used Phonopy code [31] to calculate the phonon dispersion relations for the above materials. As shown in Fig. 5 (c), there are no imaginary phonon modes (negative frequencies) indicating those three materials are dynamically stable at 0 K.
3 Discussion

In this work, we propose a physics guided deep crystal generative model (PGCGM), in which two kinds of physics based losses are invented in the generator to improve the quality of generated materials. The atom distance based losses constrain the atom distance in a certain range in the generated materials and further the generated lattice parameters follow into reasonable range too. To fulfill the symmetry requirements, the model transforms the implicit rules between base atoms sites and full atom sites into explicit cost functions. Two baseline methods are compared and PGCGM achieves the best performance across all evaluation metrics. In particular, PGCGM significantly outperforms the two baseline models in terms of property distribution metric which is a much stronger indicator to show the reality of the generated materials [35]. In addition, we use BOWSR to optimize 2000 randomly selected methods in each method. Our approach has the best match rate calculated between the Generative model-generated materials and BOWSR-optimized materials, which further demonstrate our method can generate realistic materials.

In order to see how our approach can rediscover materials in existing databases, we sample different size of materials and calculate rediscover rate for training and test datasets. We can observe a clear trend of increased rediscover rate over sampling size. There is no clear saturation point of rediscover rate at the end of 60 million
sampled materials as in CubicGAN [36]. The reasons are: 1) the possible design space of 20 space groups (5 crystal systems) in this work are much bigger than 3 space groups (only cubic crystal systems) in CubicGAN; 2) CubicGAN uses special fractional coordinates while PGCGM generates fractional coordinates in real space, which means PGCGM has a much broader space to explore new materials.

Furthermore, 1,869 out of 2,000 materials are successfully optimized by DFT calculation. Among 1,869 materials, 39.6% are with negative formation energy and 5.3% are with energy above hull less than 0.25 eV/atom, indicating that invented physics guided losses help generative crystal structures effectively. This research gives a deep insight on physics losses in the generator and provides a new way to expand the diversity of generated materials.

4 Methods

Mathematical conversion in crystal representations  Fractional coordinates can be converted to Cartesian coordinates $[x, y, z]^T$ using [9]:

$$
\begin{bmatrix}
  x \\
  y \\
  z
\end{bmatrix} = A \cdot \begin{bmatrix}
  u \\
  v \\
  w
\end{bmatrix},
$$

(7)

where $A$ is a lattice matrix calculated by lattice parameters $P$ using:

$$
A = \begin{bmatrix}
  a & b \cos \gamma & c \cos \beta \\
  0 & b \sin \gamma & \frac{c \cos \alpha - c \cos \beta \cos \gamma}{\sin \gamma} \\
  0 & 0 & \frac{V}{ab \sin \gamma}
\end{bmatrix},
$$

(8)

where $V = abc \sqrt{1 - \cos^2 \alpha - \cos^2 \beta - \cos^2 \gamma + 2 \cos \alpha \cos \beta \cos \gamma}$ is the volume of the unit cell.

In order to acquire all atom positions in the unit cell, each base atom site can be converted by affine matrix $O$. The conversion procedure is summarized in Algorithm 1. Different materials vary from the number of atoms and the number of elements. In order to make a fixed size of inputs, we only use ternary materials in this research. After conversion shown in Algorithm 1, the number of atom (sites) also differs from materials. That is the reason why base atom sites (one element one base site) are used to represent atom positions. In addition, it should be noted that the calculation of the uniqueness at line 10 of Algorithm 1 is not differentiable and time-consuming.

Algorithm 1 Generate unique coordinates using base sites and affine matrix

Require: The space group $sgp$, the base atom sites $B$
1: $O \leftarrow \text{Lib}(sgp)$ \hfill $\triangleright$ Lib saves affine matrix
2: $n \leftarrow \text{len}(O)$
3: $\text{coords} \leftarrow \text{an empty list}$
4: for $i \leftarrow 1$ to $3$ do
5: \hspace{1em} add 0 to $b_i$
6: \hspace{1em} $\text{uniq} \leftarrow \text{an empty list}$
7: \hspace{1em} for $j \leftarrow 1$ to $n$ do
8: \hspace{2em} $c \leftarrow b_i \cdot t_j - \lfloor b_i \cdot t_j \rfloor$
9: \hspace{2em} pop last element from $c$
10: \hspace{2em} if $c$ not in $\text{uniq}$ then
11: \hspace{3em} add $c$ to $\text{uniq}$
12: \hspace{2em} end if
13: \hspace{1em} end for
14: add $\text{uniq}$ to $\text{coords}$
15: end for
16: return $\text{coords}$

Materials Representation  In previous section we use $M = (E, B, P, O)$ to completely describe a crystal material. As shown in mainframe of PGCGM, however, we use three sets of base atom sites $(B^0, B^1, B^2)$. Thus here we re-formulate a material as $M^* = (B^0, B^1, B^2, P, E, sgp)$. The space group $sgp$ is used to link to the affine matrix $O$. We can use $(B^0, B^1, B^2)$ in $M^*$ to calculate physical properties as inputs to the discriminator and to design physics-based losses. Three sets of base atom sites are useful for two reasons: (1) we want to add more crystal information for the discriminator and let the discriminator have enough information to tell real
materials from fake ones; (2) With more base atom sites, we can calculate more atom distances as the physical constraints in the generator and the inputs to the discriminator.

**All Fractional Coordinates** We use affine matrix $O$ to acquire the whole atom sites in the unit cell as shown in Algorithm 1. Since the number of affine operators in $O$ varies in space groups, we zero-pad the affine matrices as large as $192 \times 4 \times 4$. We then transform each base atom site by the affine matrix and get a coordinates matrix $F_{all}$ with shape of $192 \times 3 \times 3$. Affine transformation leads to duplicate fractional coordinates. In material science, practitioners usually remove the duplicates. However, uniqueness calculation is not differentiable and it requires lots of time to do it. We choose to average along with the first dimension of $F_{all}$ to get three sets of averaged full fractional coordinates $(F^0, F^1, F^2)$, each of which is with shape of $3 \times 3$.

For a real material, base atom sites $(B^0_{\text{real}}, B^1_{\text{real}}, B^2_{\text{real}})$ can be transformed into the same average full fractional coordinates, which means $F^0_{\text{real}} = F^1_{\text{real}} = F^2_{\text{real}}$. When generating a fake material, base atom sites $(B^0_{\text{fake}}, B^1_{\text{fake}}, B^2_{\text{fake}})$ are supposed to belong to the same fake material, which hopefully results in $F^0_{\text{fake}} = F^1_{\text{fake}} = F^2_{\text{fake}}$. However, the transformation of $(B^0_{\text{fake}}, B^1_{\text{fake}}, B^2_{\text{fake}})$ might slightly deviate from the goal. Thus using $(F^0, F^1, F^2)$ in real and fake materials implicitly adds physical constraints, which somehow pushes the generator to generate different sets of base atom sites for a same material, which increases chances to generate good materials in return.

**Base Cartesian Coordinates** Three sets of Cartesian coordinates can be calculated for each set of base atom sites by Eq. (7) and we denote them by $(C^0, C^1, C^2)$.

**Atom Distance Matrices** Given three sets of base atom sites $(B^0, B^1, B^2)$, we calculate the atom distance matrices $H_{\text{inter}}$ and $H_{\text{intra}}$ as shown in sub-figure (d) of Figure 1. We firstly calculate pair-wise different atom distance matrix for each base atom site $B^j$, $j = 0, 1, 2$ and return only values in upper triangle of corresponding distance matrix termed by $H_{\text{inter}}$. Then we select three atoms belonging to the same element to form a set of three atom sites for three elements and calculate pair-wise same atom distance matrix and again return only values in upper triangle of corresponding distance matrix termed by $H_{\text{intra}}$. The final shape of $H_{\text{inter}}$ and $H_{\text{intra}}$ both is $3 \times 3$.

**Lattice Parameters** The volume of the unit cell can be calculated by lattice parameters $P$. We repeat the scalar volume three times to get the volume vector $V$. We also use the lattice matrix $A$ in Eq. (8) as part of the inputs to the discriminator.

**Element Properties** We use 23 properties as shown in Table 6 to formalize element matrix $E$.

| Properties | Properties | Properties |
|------------|------------|------------|
| Atomic number | Average ionic radius | noble gas or not |
| Pauling electronegativity | Average cationic radius | transition metal or not |
| Periodic table row | Average anionic radius | post transition metal or not |
| Periodic table group | Sum of all ionic radii | metalloid or not |
| Atomic mass | Maximum oxidation state | alkali or not |
| Atomic radius | Minimum oxidation state | alkaline or not |
| Mendeleev number | Average all common oxidation states | halogen or not |
| Molar volume | Average all known oxidation states | |

Now we list all parts of inputs to the discriminator in Table 7. $P^*$ only contains the lengths because the angles are either $(90^\circ, 90^\circ, 90^\circ)$ or $(90^\circ, 90^\circ, 120^\circ)$ in the training materials. Thus instead of generating three angles in $P$ for fake materials, we build a map between angles and the space group $sgp$. Then we concatenate all parts and a zero matrix of shape $3 \times 3$ into a matrix of shape of $3 \times 64$. The matrix is finally reshaped into $3 \times 8 \times 8$ as the inputs to the discriminator.

| symbol | shape | symbol | shape |
|--------|-------|--------|-------|
| $(B^0, B^1, B^2)$ | $3 \times 9$ | $P^*$ | $3 \times 1$ |
| $(F^0, F^1, F^2)$ | $3 \times 9$ | $V$ | $3 \times 1$ |
| $(C^0, C^1, C^2)$ | $3 \times 9$ | $(H_{\text{inter}}, H_{\text{intra}})$ | $3 \times 6$ |
| $E$ | $3 \times 23$ | $A$ | $3 \times 3$ |

**Discriminator** There are two input branches for crystal representation and affine matrix in Discriminator as in sub-figure (b) of Figure 1. Each branch is forwarded to a 2D convolutional block and the learnt features are con-
catenated together. The concatenated vector is sent to a couple of fully connected layers to get the discriminative score. We have three different sets of base atom sites in our inputs and with the affine matrix branch, it helps to implicitly learn the knowledge of how affine matrix transforms base atom sites into full atom sites. The detailed architectures of two convolutional blocks can be found in Table S3 in the supplementary materials.

**Generator**  The architecture of generator is shown in sub-figure (a) of Figure 1. Three branches are found. Conditioning on element constituents and space group, the generator outputs three sets of base atom sites \((B_{fake}^0, B_{fake}^1, B_{fake}^2)\) and unit cell length \(P^*\). Then we re-formalize Eq. (2) as follow:

\[
(B_{fake}^0, B_{fake}^1, B_{fake}^2, P^*) = f_\theta^* (Z, E, sgp).
\]  

Taking random noise \(Z\), space group \(sgp\), and element properties matrix \(E\) as inputs, the generator can generate a material with the same lattice parameters and space group but different representations of the base atom sites when merely sampling one material. Our goal here is that the generated three sets of base atom sites belong to the same material. Random noise \(Z\) is mapped to a dense vector a fully connected layer. The space group branch is the same as in discriminator. Element matrix \(E\) is forwarded to a 1D convolutional layer (Conv1D). The outputs of random noise and space group branches are combined together as the inputs to a multi-layer perceptron (MLP) block to generate unit cell length \(P^*_{fake}\). The outputs of random noise and element branches are combined together as the inputs to 2D deconvolutional layers (ConvTran2D) to generate three sets of base atom sites \((B_{fake}^0, B_{fake}^1, B_{fake}^2)\). The detailed descriptions for MLP, Conv1D, and ConvTran2D can be found in Table S4 in the supplementary materials.

**Atom Clustering and Merging**  For crystals with high symmetry, the number of atoms in the unit cell tends to be very large after conversion by Algorithm 1. We propose a post-processing method to reduce the number of atoms by clustering and merging. Firstly, we cluster the nearby atoms of the same elements by forming flat clusters from hierarchical clustering [2, 23]. The maximum atom distance allowed in our research is 1.2 times the atom radius sum. Secondly, we merge the atoms in the same clusters considering periodic attributes of crystal structures.

**Evaluation Metrics**  Past studies in crystal generation used different evaluation metrics, making it hard to compare different methods. Here, we create a set of metrics to evaluate our method and two baselines. 1) **Validity.** Following [5], we consider a crystal structure as valid when the shortest distance between any two atoms is bigger than 0.5Å. Following CubicGAN, we calculate the overall charge of a crystal structure using pymatgen [26] and if it is neutral, then it is valid. Also, we count the number of structures after post-processing in our method and we apply the same post-processing onto the CubicGAN. 2) **Property distribution.** We calculate wasserstein distance (WD) between the property distribution of generated materials and materials in test dataset TST. The properties we used are minimum atom distance, maximum atom distance, and density. 3) **Diversity.** We calculate the diversity of compositions, which means the ratio of unique number of compositions in generated structures.

**References**

[1] Maximilian Amsler and Stefan Goedecker. Crystal structure prediction using the minima hopping method. *The Journal of chemical physics*, 133(22):224104, 2010.

[2] Ziv Bar-Joseph, David K Gifford, and Tommi S Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1):S22–S29, 2001.

[3] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science*, 58(3):364–369, 2002.

[4] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[5] Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of chemical information and modeling*, 60(10):4518–4535, 2020.
[6] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):1–7, 2020.

[7] K Doll, JC Schön, and M Jansen. Structure prediction based on ab initio simulated annealing for boron nitride. *Physical Review B*, 78(14):144110, 2008.

[8] José A Flores-Livas. Crystal structure prediction of magnetic materials. *Journal of Physics: Condensed Matter*, 32(29):294002, 2020.

[9] Fractional coordinates. Fractional coordinates — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/wiki/Fractional_coordinates#cite_note-3. [Online; accessed 11-November-2021].

[10] Alberto Franceschetti and Alex Zunger. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature*, 402(6757):60–63, 1999.

[11] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[14] Richard Hill. The elastic behaviour of a crystalline aggregate. *Proceedings of the Physical Society. Section A*, 65(5):349, 1952.

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

[16] Jordan Hoffmann, Louis Maestrati, Yoshihide Sawada, Jian Tang, Jean Michel Sellier, and Yoshua Bengio. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv preprint arXiv:1909.00949*, 2019.

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

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

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

[20] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

[21] Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Guttfleisch, and Hongbin Zhang. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):1–7, 2021.

[22] Félix Mouhat and François-Xavier Coudert. Necessary and sufficient elastic stability conditions in various crystal systems. *Physical review B*, 90(22):224104, 2014.

[23] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[24] Juhwan Noh, Jae hoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and You sung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1 (5):1370–1384, 2019.
[25] Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. Crystalgan: Learning to discover crystallographic structures with generative adversarial networks. In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, 2019.

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

[27] Warren E. Pickett and Jagadeesh S. Moodera. Half metallic magnets. Physics Today, 54(5):39–44, 2001. doi:10.1063/1.1381101. https://doi.org/10.1063/1.1381101. URL https://doi.org/10.1063/1.1381101.

[28] Edward O Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. Annual Review of Materials Research, 45:195–216, 2015.

[29] Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. arXiv preprint arXiv:2005.07609, 2020.

[30] Yoshiihide Sawada, Koji Morikawa, and Mikiya Fujii. Study of deep generative models for inorganic chemical compositions. arXiv preprint arXiv:1910.11499, 2019.

[31] A Togo and I Tanaka. First principles phonon calculations in materials science. Scr. Mater., 108:1–5, Nov 2015.

[32] Vei Wang, Nan Xu, Jin-Cheng Liu, Gang Tang, and Wen-Tong Geng. Vaspkit: A user-friendly interface facilitating high-throughput computing and analysis using vasp code. Computer Physics Communications, 267:108033, Oct 2021. ISSN 0010-4655. doi:10.1016/j.cpc.2021.108033. URL http://dx.doi.org/10.1016/j.cpc.2021.108033.

[33] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Calypso: A method for crystal structure prediction. Computer Physics Communications, 183(10):2063–2070, 2012.

[34] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters, 120(14):145301, 2018.

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

[36] Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. arXiv preprint arXiv:2102.01880, 2021.

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[38] Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with bayesian optimization and graph deep learning. Materials Today, 51:126–135, 2021.