

8-5-2021

High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks

Yong Zhao

Mohammed Al-Fahdi

Ming Hu

Edirisuriya M.D. Siriwardane

Yuqi Song

See next page for additional authors

Follow this and additional works at: https://scholarcommons.sc.edu/csce_facpub



Part of the [Computer Sciences Commons](#), and the [Engineering Commons](#)

Publication Info

Published in *Advanced Science*, Volume 8, Issue 20, 2021, pages 2100566-.

© 2021 The Authors. *Advanced Science* published by Wiley-VCH GmbH

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

This Article is brought to you by the Computer Science and Engineering, Department of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

Author(s)

Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya M.D. Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu

High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks

Yong Zhao, Mohammed Al-Fahdi, Ming Hu,* Edirisuriya M. D. Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu*

High-throughput screening has become one of the major strategies for the discovery of novel functional materials. However, its effectiveness is severely limited by the lack of sufficient and diverse materials in current materials repositories such as the open quantum materials database (OQMD). Recent progress in deep learning have enabled generative strategies that learn implicit chemical rules for creating hypothetical materials with new compositions and structures. However, current materials generative models have difficulty in generating structurally diverse, chemically valid, and stable materials. Here we propose CubicGAN, a generative adversarial network (GAN) based deep neural network model for large scale generative design of novel cubic materials. When trained on 375 749 ternary materials from the OQMD database, the authors show that the model is able to not only rediscover most of the currently known cubic materials but also generate hypothetical materials of new structure prototypes. A total of 506 such materials have been verified by phonon dispersion calculation. Considering the importance of cubic materials in wide applications such as solar panels, the GAN model provides a promising approach to significantly expand existing materials repositories, enabling the discovery of new functional materials via screening. The new crystal structures discovered are freely accessible at www.carolinamatdb.org.

1. Introduction

Data-driven accelerated design of new materials is emerging as one of the most promising approaches for addressing the challenges in finding next-generation materials. Currently, one of the main strategies for materials discovery is screening existing materials databases.^[1–4] However, such approaches are severely limited by the scale and diversity of the existing structures in the repositories, such as Inorganic Crystal Structure Database (ICSD) and Materials Project (MP), which have about 165 000 and 125 000 materials, respectively, compared to the almost infinite chemical design space. For example, lithium compounds are widely used in electric vehicles and mobile phone batteries, but there are only 16 000 different lithium compounds in the MP database, which has been almost exhaustively screened for better lithium-ion battery.^[5,6] Large-scale generation of stable hypothetical crystal structures is strongly needed to significantly expand the current materials repositories in both the quantity and compositional and structural diversity


to increase the success rate of high-throughput screening of novel functional materials.

The properties of materials are closely linked to their crystal structures. Traditionally, materials scientists discover new materials by either trial-and-error or heuristic random-guess approaches, both of which are notoriously labor-intensive. One example database is ICSD,^[7] which collects almost all discovered materials since 1913. To date, only around 165 000 experimental structures are reported in ICSD. Considering the number of elements in the periodic table and their possible combinations, the design space of materials would be infinite combinatorially. Hence, better approaches for new materials discovery are needed.

Several working directions are investigated for the generation of new materials.^[8–16] There are mainly three different ways to generate or discover new crystal structures including doping/element substitution,^[5,17–19] composition generation plus crystal structure prediction,^[9] and generative machine learning models.^[12,14–16,20,21] The element substitution approach is the most widely used strategy. But it is subject to the extremely limited known prototype structures in the database compared to the

Y. Zhao, E. M. D. Siriwardane, Y. Song, A. Nasiri, J. Hu
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201, USA
E-mail: jianjunh@cse.sc.edu

M. Al-Fahdi, M. Hu
Department of Mechanical Engineering
University of South Carolina
Columbia, SC 29201, USA
E-mail: hu@sc.edu

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.202100566>

© 2021 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202100566

vast chemical design space. The second approach can exploit the recently developed generative models^[12] to generate a large number of hypothetical materials compositions and then use crystal structure prediction codes to predict their structures. Many global optimization methods have been developed to search the appropriate compositions and structures, including simulated annealing,^[22] basin hopping,^[23] minima hopping,^[24] genetic and evolutionary algorithms.^[10,11] Those approaches generally guide the searches toward the local minima of free energy to identify the stable or meta-stable structures either by initial configuration space or chemical composition. However, these crystal structure prediction algorithms are usually too computationally expensive due to their reliance on density functional theory (DFT)-based formation energy calculation and can thus only handle relatively simple structures. For complex structures, most of the time, these methods fail to find the ground truth structures corresponding to the global minimum formation energy.

One of the most promising approaches for new materials structure creation is deep generative machine learning models.^[12–16,25–27] Both variational autoencoder (VAE)^[14,15,26,27] and generative adversarial networks (GAN)^[12,13,16,25] have been adapted for inverse design of inorganic materials with different crystal structure representations. A VAE model contains two parts: an encoder and a decoder.^[28–30] The encoder part encodes the crystal structure distribution into a latent space, and the decoder reconstructs the material structures from the latent space. After training, new material structures can be generated by sampling in the latent space. Conversely, a GAN generator model consists of two neural networks: a discriminator (critic) and a generator, both of which are trained simultaneously. The discriminator is trained to differentiate real materials from fake ones generated by the generator, while the generator tries to generate fake materials as real as possible to fool the discriminator. The Nash equilibrium achieved by the discriminator and the generator helps a GAN learn the distribution of the materials implicitly. In the past few years, several inorganic materials generative models have been proposed. Those works are limited by their chemical family (e.g., special oxides)^[13,15,16] or formulas generation^[12] or hydrides.^[25] Noh et al.^[15] present a framework for learning a continuous vector for vanadium oxides using VAE, which is trained on a 3D image-like representation to attain the continuous materials space. Two sampling strategies are used in the latent space to generate only V_xO_y materials. Training the VAE model using 3D grid representation is computationally demanding and memory-hungry. In, ref. [16] Kim and Noh et al. trained a composition family-specific GAN model on the Mg-Mn-O system using the atom coordinates as the representation of materials. The crystal GAN model is composed of three modules: a generator, a critic (discriminator), and a classifier. The critic calculates the Wasserstein distance between real and fake materials.^[31] The classifier module ensures that the generator generates desired composition and atom numbers in the unit cell. However, this model can only be used to generate structures of the Mg-Mn-O system, and the model quality is limited by the small dataset since there are only limited known compounds of this chemical system. CrystalGAN,^[25] proposed by Nouria et al., consists of a cross-domain GAN model, which maps one hydride system into another using CycleGAN schema.^[32] All these works focus on generating materials of a special material system. In a

most recent work, Ren et al.^[14] proposed a new VAE model that directly uses the atom coordinates and unit cell lattice parameters to encode the structures. To constrain the neural network model behavior, their invertible representation encodes the crystallographic information into the descriptors in both real space and reciprocal Fourier space crystal properties. Their model is trained with 24 785 unique ternary materials and can generate interesting new structures. However, most of their new structures are generated by perturbing the latent vectors of known materials. Large-scale generation of stable crystal structures remains a challenging problem. Other than generating materials structures, Dan et al.^[12] proposed MatGAN to generate millions of novel material formulas with chemical validity, which expands the candidates for inverse design of new solid materials.

In this work, we propose a novel deep generative model called CubicGAN to generate cubic materials structures on a large scale. Ternary materials selected from the open quantum materials database (OQMD)^[33] database are chosen as our training set because of its large size of materials and diverse compositions. In our model, material structures are represented by their lattice parameters, atom coordinates, element embedding, and the space group. The conditions of a specific space group and three elements are fed to the generator to generate desired crystal material structures. We trained ternary and a quaternary GAN models to generate novel cubic (ternary and quaternary) crystal structures of the space groups 216, 255, 221. Materials of these three space groups consist of 78.5% of all ternary and quaternary cubic materials in OQMD, covering a majority of known cubic materials space.

Our systematic experiments show that our CubicGAN model can recover not only many of the known cubic structures but also discover many new materials with new composition prototypes with different anonymous formulas (new prototypes). Additional large-scale DFT based validation has led to the discovery of 506 new cubic crystal materials of new prototypes. The detail of the CubicGAN model will be explained in the following sections. Compared to refs. [15,16,25] our framework can generate a large variety of materials of different chemical systems. The only work that is similar to ours in terms of variety of materials is ref. [14] in which Ren et al. use VAE rather than GAN as the generative model trained with train ternary materials in Materials Project^[34] database. However, their model tends to generate new samples by interpolation. The second major difference is a much simpler representation used in our work without the momentum space representations.

Our contributions can be summarized as follows:

- i) We propose a novel GAN model to generate large-scale cubic materials conditioning on the elements and a specified spacegroup. In total, we generate 10 million hypothetical ternary and 10 million quaternary crystal structures for downstream analysis.
- ii) We perform three stage checks on generated materials and extensively match the generated materials against existing databases. The results show that our method can rediscover a majority of cubic materials in the existing databases. In addition, most of the rediscovered materials from MP are confirmed as stable or meta-stable materials in terms of energy-above-hull.

Table 1. Statistics of OQMD ternary and quaternary materials (Total 813 839).

	Ternary	Ternary cubic	Ternary with three nonequivalent sites	Ternary cubic with 3 nonequivalent sites
Count	556 839	505 456	411 646	388 680
Cubic percentage	505456/556839 = 91%		388 680/411 646 = 94%	
	Quaternary	Quaternary cubic	Quaternary with four nonequivalent sites	Quaternary cubic with 4 nonequivalent sites
Count	141 100	127 659	129 514	127 523
Cubic percentage	127659/141100 = 90%		127 523/129 514 = 98%	

- iii) We perform DFT simulations on 108 897 hypothetical materials, of which 33.8% novel materials are successfully relaxed. By further analysis, we demonstrate that new crystal structure prototypes (with different anonymized formula types) can be found, such as ABC_6 –216, ABC_6D_6 –216, and AB_8C_{12} –221.
- iv) By further stability verification, 506 new-prototype materials have been generated and confirmed to be stable by phonon dispersion calculation.

2. Experimental Section

In this work, the authors focused on training generators of ternary and quaternary cubic crystal structures of three space groups (216, 221, 225) to simplify this model design while ensuring coverage of a majority of the cubic design space. It was found that in the OQMD dataset with 813 839 materials, 85.8% of them were ternary or quaternary materials. In addition, out of all the cubic crystals, 97.8% of them belonged to these three space groups, again covering the majority of the known cubic materials space. These three space groups were selected because it was found that for the materials of these three space groups, most of their nonequivalent atom fractional coordinates in the CIF files had a multiplicative factor of 0.25 or belong to this set [0, 0.25, 0.5, 0.75]. So, instead of generating cubic structures with arbitrary real-valued atom coordinates, it was only aimed to train a cubic material generator that only generated structures whose atom positions were sitting at positions with their fractional coordinate values to be from this set $\pm 0, 0.25, 0.5, 0.75$. In this case, the special discrete fractional coordinates were much easier to be generated accurately by these deep neural networks. This decision had dramatically simplified the generation model, and thus the authors chose the training data with these two criteria: ternary and quaternary cubic crystal structures of three space groups (216, 221, 225).

2.1. Dataset

The training data from OQMD,^[33,35] which is an open-source database of experimental and DFT-calculated materials were collected. Totally 813 839 entries were retrieved from version 1.3 of OQMD. Entries calculated with local-density approximation (LDA) were also included. Among them, 556 839 and 141 100 POSCAR files were successfully build for ternary and quaternary materials in the OQMD, of which 505 456 and 127 659

structures belonged to cubic crystal systems, respectively. After converting the POSCAR files to symmetrized CIF files, 411 646 ternary materials had three unique nonequivalent atom sites, of which 388 680 materials of cubic crystal systems were found; 129 514 quaternary materials had four unique nonequivalent atom sites, of which 127 523 materials belonged to the cubic crystal systems. **Table 1** shows the statistics of OQMD materials distributions. It was found that ternary materials of cubic crystal systems were the largest chunk (91%) out of all ternary materials. Similarly, it was observed that the ternary cubic structures with three nonequivalent sites were 94% out of all ternary materials with three nonequivalent sites. For quaternary materials, these two percentages were 90% and 98%, respectively. This means that this CubicGAN model could be used to generate hypothetical cubic materials that are the majority type of known material category.

Another key criterion for selecting the training samples was that cubic structures with three nonequivalent atom positions (in CIF files) were only picked for training ternary GAN model (for quaternary GAN, the number is four). Making this choice allowed the authors to use a unified matrix of dimension (28×3) to represent all ternary cubic materials (for quaternary materials, the dimension is 27×3 where only one space group is used in this work). For a given material, once its nonequivalent positions and space group are determined, the full atom positions within the unit cell can be calculated by symmetry operations. 411 646 ternary materials had been identified with only three nonequivalent positions, of which 388 680 (94%) materials belonged to cubic crystal systems as shown in Table 1. Out of these 388 680 materials, 22 space groups were found as shown in Figure S1, Supporting Information. Among them, the space groups that had the most numbers of materials were $Fm\bar{3}m$ and $F\bar{4}3m$ (the total portion of these two space groups was 97.2%). $Pm\bar{3}m$ was the third one with only 6462 samples or 1.7%. After removing the duplicate cubic materials within MP and ICSD, almost all the materials (375 749 out of 384 215) with the space groups of $Fm\bar{3}m$, $F\bar{4}3m$, and $Pm\bar{3}m$ followed this criterion.

Table 2 shows the overall statistics of the finalized training and validation datasets. In total, 375 749 ternary materials were selected from three cubic system space groups from OQMD to form the OQMD-TC3 (T: Ternary, C: Cubic, 3: three space groups) training dataset: $Fm\bar{3}m$, $F\bar{4}3m$, and $Pm\bar{3}m$ each having 186 344, 184 162, and 5243 materials, respectively. These materials together corresponded to 249 646 unique formulas. With this diversity of formulas, the CubicGAN model could efficiently learn valid combinations of ternary elements. The unique 84 elements

Table 2. Statistics of our training data and validation datasets from OQMD, MP, and ICSD. Here the cubic materials (6545 + 1875) from MP/ICSD are used as validation set and are excluded from the training set.

Ternary Materials						
Dataset	Total	Fm $\bar{3}$ m	F $\bar{4}$ 3m	Pm $\bar{3}$ m	Unique formula	Unique element
Training:OQMD-TC3	375 749	186 344	184 162	5243	249 646	84
Validation:MP-TC3	6545	4576	520	1449	6431	84
Validation:ICSD-TC3	1875	804	280	791	1034	84

Quaternary Materials			
Dataset	Total	Unique formula	Unique element
Training:OQMD-QC1	121 008	39 767	56
Validation:MP-QC1	39	39	39
Validation:ICSD-QC1	8	7	12

in the datasets were utilized to generate random three-element combinations during GAN training. The same steps were applied to quaternary materials in OQMD. As shown in Figure S1, Supporting Information, materials with space group F $\bar{4}$ 3m occupies 95% of the quaternary data. So for training the quaternary GAN model, materials of space group F $\bar{4}$ 3m were only chosen.

The ternary data from the Materials Project and ICSD databases were used as validation sets to check the rediscovery rates for the proposed method. The ternary materials in Materials Project database^[34] and ICSD^[7] were processed as it was done for OQMD samples to create the MP-TC3 and ICSD-TC3 validation datasets. In total, 6545 cubic materials were retrieved, of which the numbers of materials with Fm $\bar{3}$ m, F $\bar{4}$ 3m, and Pm $\bar{3}$ m were 4576, 520, and 1449, respectively and there were 6431 unique formulas existing in the whole retrieved data. From the ICSD database, 1875 cubic materials were found to satisfy the selection criteria, of which the numbers of materials were 804, 280, and 791 for space groups Fm $\bar{3}$ m, F $\bar{4}$ 3m, and Pm $\bar{3}$ m. For quaternary materials, the OQMD-QC1 training dataset had 121 008 samples. However, only 39 and 8 quaternary materials were found in MP and ICSD that satisfy the two selection criteria (see Table 2). These samples were removed from the training dataset selected from OQMD by removing the crystal structures with a minor difference of cube lengths from the samples in the validation sets).

In terms of prototypes in the validation datasets MP-TC3 and ICSD-TC3, Table S3, Supporting Information, shows details of the existing prototypes for materials that satisfy the selection criteria. The prototype “ABC2-225” was taken as an example. Here ABC2 and 225 were the crystal prototype anonymous formula and the space group number used to denote a prototype, and this format was used in the following content. Overall, the three databases had the same set of prototypes; other than that, MP had an extra one: AB6C6-225. However, only one material (mp-1147668) was found under AB6C6-225 and was unstable. For quaternary materials in OQMD, there were only two prototypes, including ABCD-216, with 121 006 materials and ABCD₆-216 with two materials. Moreover, it was found that quaternary cubic materials distribution was highly biased with 121 018 belonging

to space group 216, and only 5674 belonging to space group 225, and no samples found for space group 221. For simplicity, the quaternary CubicGAN was trained using only the samples from space group 216 and it then could only generate samples of this space group.

2.2. CubicGAN Framework

Figure 1 illustrates the main framework of this method. The framework primarily contained two steps: GAN training and material generation. The goal was to train a generator that learns the distribution from known materials data and then sample from it. To achieve this, the generator was trained to create fake material structures, conditioned on a given space group and a specification of three elements. The three elements were randomly chosen from 84 elements in the dataset. The 84 elements were one-hot encoded and were converted to a 3×23 element matrix by the embedding layer. There are several ways to encode elements such as learnt atom embedding,^[36,37] atom2vec encoding,^[38] atom embedding learnt from research literature,^[39] and primitive element properties. In this work, the authors chose to use primitive elemental properties as the element embedding approach and use them to set the weights in the element embedding layer. The reason was that CubicGAN learnt to build valid crystal structures from three elements and their properties, coordinates, and space-group. With raw elemental properties, it was beneficial for CubicGAN to find the most desirable combinations of elements, coordinates, and space groups. The parameters of the embedding layer were initialized by 23 element properties as shown in Table S1, Supporting Information. Taking a randomly selected space group (one-hot encoded), 3-element combinations (one-hot encoded), and random noise Z as inputs, the generator then generated material structures with the specified space group and element constituents. Space groups and elements were mapped into dense vectors by their corresponding embedding layers. The number of atoms for each element did not need to be specified as it can be determined by the space group symmetry operations. The random selections of space groups were based on the portions of three cubic space groups considered in the model: Fm $\bar{3}$ m, F $\bar{4}$ 3m, and Pm $\bar{3}$ m. The detailed architecture is shown in Figure S2, Supporting Information.

An input to the discriminator has four parts: nonequivalent atomic coordinates, element properties, unit cell parameters, and space groups as shown in Figure 1. The coordinates part included the fractional coordinates of three nonequivalent atoms. For three unique elements in each material, each element was represented by 23 properties as shown in Table S1, Supporting Information. Since the lattice lengths a , b , c were the same in cubic crystals, only one value was needed to represent it. Three cubic space groups were one-hot encoded. As shown in Figure S2, Supporting Information, four parts were concatenated together to form a tensor with the dimension of 3×28 . The input was then forwarded to four 1D convolutional layers, of which the kernel size was 1×1 , which was used to capture the implicit relationships among the four parts. Two CNN layers were used to reduce the dimension from three to one. Then, a few fully connected layers were used to map them to Wasserstein estimation.^[31] The detailed network settings are shown in Figure S2,

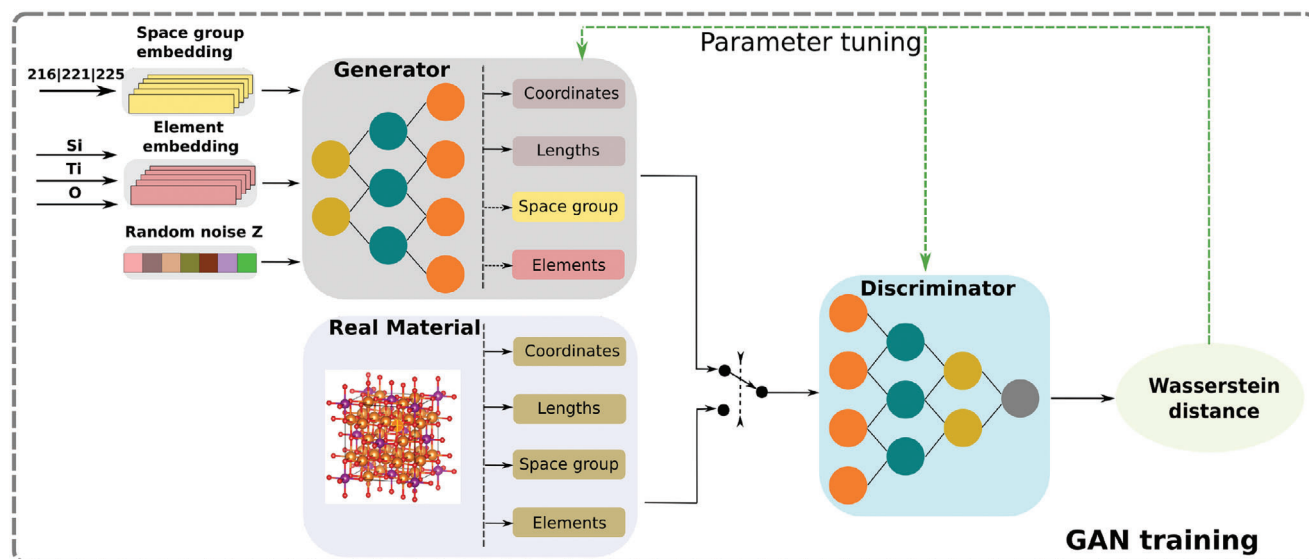


Figure 1. The workflow of our CubicGAN framework.

Supporting Information. In standard conditional GAN, the input of the generator included the random noise and a condition vector.^[40] Here, a space group embedding layer and an element embedding layer was added as shown in Figure 1 to map the randomly selected one-hot encoded space group (chosen from 216/221/225) and three randomly selected elements (one-hot encoded) into the latent vectors. The reasons for this design were as follows: 1) As only three dominant cubic space groups were used in this work, the combination of atom positions with corresponding elements, unit cell lengths, and one-hot encoded space group symmetry was sufficient to describe a material structure; 2) Using element properties as part of the representation made the generator learn to generate chemically valid materials, for example, structures that do not violate Pauling's rules. As the authors previous work^[12] shows, the composition constraints could be learned from the compositions of existing materials. Here, the CubicGAN was also configured to learn both implicit compositional as well as structural constraints to help the generator generate only valid ternary or quaternary formulas as much as possible; 3) The 2D representations of the cubic structures also matched well with the convolutional layers used in the discriminator, in which the convolutional operations could extract implicit relationships among four parts of information.

The generator and the discriminator of the CubicGAN model were trained with the loss function of Wasserstein distance^[31] which measured the dissimilarity between distribution differences of real and fake materials. Compared to loss functions used in traditional GAN,^[41] Wasserstein distance improved the model stability and prevented the mode collapse. The gradient penalty was used to clip weights in order to improve the stability of training as done by Gulrajani et al.^[42] The penalty of gradient norm with respect to the inputs worked as a regularization term to stabilize the training process of the GAN. More formally, the cost function for GAN training is as follows:

$$L = \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{\mathbf{x}}}} [D(\tilde{\mathbf{x}})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [D(\mathbf{x})] + \lambda \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_{\tilde{\mathbf{x}}}} [(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2] \quad (1)$$

where D is the discriminator, $\mathbb{P}_{\tilde{\mathbf{x}}}$ is the distribution of interpolated samples between the distribution of real materials $\mathbb{P}_{\mathbf{x}}$, and the distribution of generated materials $\mathbb{P}_{\tilde{\mathbf{x}}}$. λ is the balancing parameter, which is set to 10 in this work.

After inspecting the generated structures by the GAN, it was found that the generated lattice parameter a was often not good enough, leading to overlapping atom clusters. To address this issue, an additional post-processing step introduced to predict the lattice length a using a composition based machine learning model that the authors recently developed,^[43] which achieved a R^2 score of 0.979 for cubic lattice a prediction.

During training the GAN, real materials were randomly picked in batches. With the fused matrix of generated materials as shown in Figure 1, they were fed to the discriminator in a mixed manner. The number of iterations of the discriminator per generator iteration was set as five. The GAN model was developed using the open-source libraries of TensorFlow^[44] and Keras.^[45] More details regarding model architecture and hyper-parameter setting can be found in Table S2, Supporting Information.

3. Results and Discussion

3.1. Performance Evaluation of CubicGAN: Validity, Uniqueness, and Rediscovery Rate Analysis

There are three major criteria for evaluating generative models, namely, validity, diversity, and uniqueness.^[46] After training the ternary CubicGAN using the OQMD-TC3 dataset, we generate 10 million cubic structures of the specified three cubic space groups (225 216 221). The proportions of the samplings are set as identical to the training set, which is 49.6%–49.0%–1.4%, respectively. To evaluate the generation performance, we first check how the percentage of the generated charge-neutral samples changes with respect to the total number of generated samples. The charge neutrality check is based on Pymatgen^[47] using the common valence values of elements as defined in Pymatgen. As shown in Figure 2, the charge-neutral samples' percentage maintains

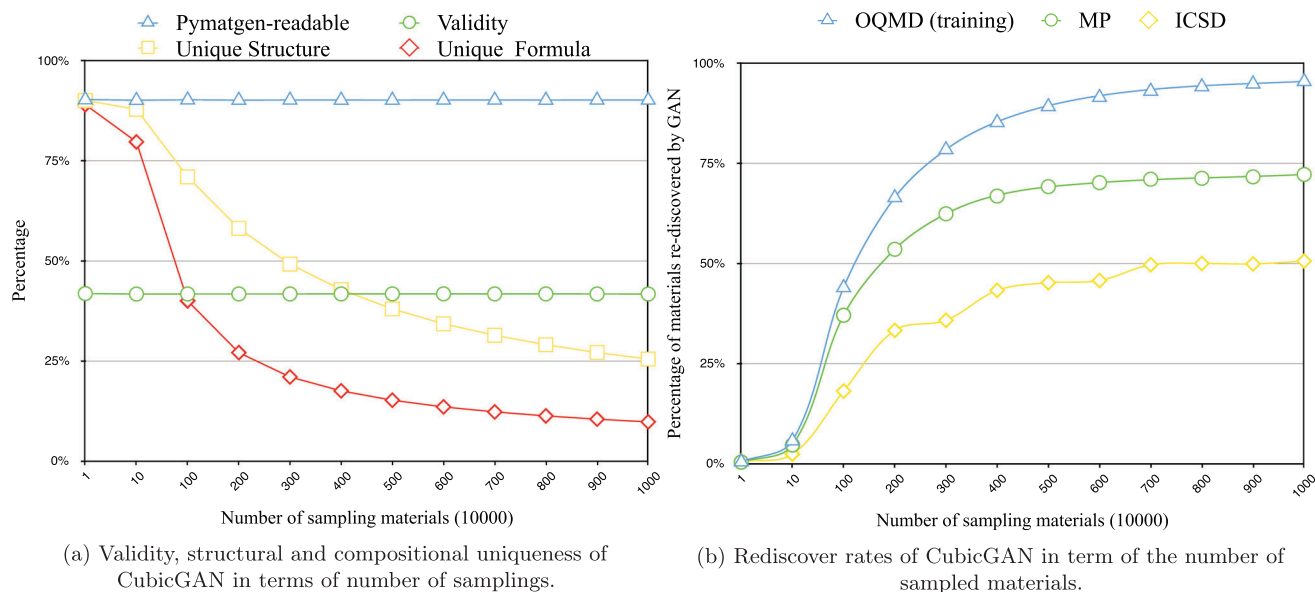


Figure 2. Performance evaluation of CubicGAN.

around 41% over the whole process of generating 10 million samples, which means that when we generate 10 million samples, appropriately we can get 4.1 million charge-neutral samples for downstream screening. We then checked how the percentage of the generated samples has pymatgen-readable crystallographic information file (CIFs), unique CIFs, and unique formulas, which reflect the diversity and uniqueness of the generator. In Figure 2, the blue line demonstrates the percentage of CIFs readable by pymatgen in terms of sampling size, and the sampling size is from ten thousand to ten million. In this work, pymatgen-readable means that CIFs can be recognized as the space group that is assigned to. We can find that the percentage of readable CIF files is stable no matter how we run the sampling. After removing the duplicate materials, we calculate the percentage of unique CIFs and unique formulas as denoted by the yellow and red lines in Figure 2. Only those materials that have the same formula and the same corresponding atom positions are considered as duplicates here. It is found that the percentages of the unique CIFs and unique formulas are decreasing and growing flat. From these observations, we believe that our GAN model might have explored the majority of the cubic crystal structure space but have not exhausted it yet.

Another effective way to evaluate the CubicGAN's performance is to check how soon it can rediscover the known cubic crystals in leave-out datasets of existing databases. To do this, in our training dataset, we have removed all the materials of the three cubic space groups (216, 225, 221) existing in MP and ICSD databases, which are 6545 and 1875, respectively. It is interesting to see how many of those leave-out cubic materials can be rediscovered by our GAN model as the sampling size goes from ten thousand to ten million. Figure 2 shows how the rediscovery percentages of the cubic crystals of the three space groups (216, 225, 221) change as the sampling size increases.

Figure 2 shows the rediscovery rates over time of sampling. At first, we check how the percentage of the rediscovered cubic samples out of all training samples (blue line) changes while

generating more samples. It is found that this training set rediscovery rate increases consistently over the sampling process. It soars quickly to 88% when the sampling size increases until 5 million samplings are reached. At the end of 10 million samplings, the rediscovery rate reaches 95.5%. Similar patterns can be observed for the rediscovery rate curve for the MP-TC3 validation dataset, as shown by the green line. With the increasing number of samplings, the rediscovery rate reaches 72.0%. This saturated percentage is much lower than that of the training set, which is due to MP-TC3 data has different proportions over the three space groups (225 216 221), which are 69.9%–7.9%–22.1%, respectively compared to 49.6%–49.0%–1.4% of the training set. Since our generation process is based on the space group proportions of the training set (which focuses on generating candidates of space groups 225 and 216, the 72% rediscovery rate is close to the percentage of these two types of samples in MP-TC3 (69.9% + 7.9% = 77.8%). We also find that half of the rediscovered materials in MP-TC3 are stable based on the formation energy and e-above-hull criteria. Details of the stability analysis can be found in Figure S3, Supporting Information. The rediscovery rate pattern over ICSD-TC3 is similar to that of MP-TC3 except that the highest rediscovery rate is 50.7% at the sampling size of ten million, which is close to the percentage of total samples of space groups 225 and 216 (42.9% + 14.9% = 56.8%). These high rediscovery rates over the training set and the two validation sets demonstrate that our CubicGAN has learned the implicit chemical rules of the cubic structures to generate in a much better way than random sampling. After the sampling size reaches 7 million, the number of materials rediscovered converges, indicating that ten million samplings could be a reasonable size to cover most of the cubic structures since they seem to have almost exhaustively explored the search space of materials that meet our criteria. Therefore, we use ten million samplings for further analysis.

To compare how our CubicGAN performs compared to random sampling or exhaustive enumeration, we calculate the enrichment score for our ternary CubicGAN. As we are searching

Table 3. Statistics of generated materials.

	Valid CIFs	Unique formulas	Crystal prototypes
Ternary	2 558 678	990 319	31 (24)
Quaternary	5 498 267	1 797 592	3 (1)
No lanthanoid and actinoid			
Ternary	1 064 650	403 337	31 (24)
Quaternary	4 382 130	1 431 500	3 (1)

candidates of three cubic space groups with three unique sites of three distinct elements and the only possible fraction coordinates are 0 0.25 0.5 0.75, the total possibility of configurations are $(4^3)^3 \times 85 \times 84 \times 83 \times 3 = 466055331840$, which is much larger than the corresponding combinations of the ternary composition space.^[12] Considering that with 10 million samplings, we have rediscovered 95.5% of the OQMD-TC3 dataset, the enrichment score is $\approx 44\,507$, which is a significant boosting for generating chemically valid crystal structures compared to exhaustive enumeration.

3.2. Large Scale Generation of New Cubic Crystal Materials Structures

While rediscovery rate analysis over the MP-TC3 and ICSD-TC3 validation sets have demonstrated the accelerated sampling in cubic structure space, there are only $6545 + 1875 = 8420$ validation samples plus the 358 840 rediscovered training samples. It is still desirable to check the chemical validity of the remaining 96.33% generated samples and filter out those promising new materials. With 10 million hypothetical cubic materials, it is impractical to perform DFT calculations for all of them to verify their chemical validity and stability. Here we adopt three stages of validation check to reduce the pool of samples for DFT validation. We use the CGCNN based graph neural network model for formation energy prediction, which was trained with samples from Materials Project database.^[36] Then we scan the generated materials in the order of space group match, charge neutrality, and formation energy filtering. The nonequivalent coordinates are transformed by symmetry operations provided by relevant space groups used when generating samples. With the full coordinates set, elements, unit cell parameters (unit cell length a and angles, which are always 90° in cubic systems), we could write a CIF. The space group check is performed by Pymatgen^[47] in the first place (we refer to this check as a Pymatgen-recognizable check). If the generated sample cannot be recognized by Pymatgen or the space group analyzed by Pymatgen is not consistent with the space group given to the generated sample, this sample is considered as a failed generated case. As shown in Table 3, in total there are 2 558 678 and 5 498 267 valid ternary and quaternary CIFs have been found from 10 million generations, respectively. From them, candidate materials with charge neutrality and CGCNN-predicted negative formation energy are reserved for further DFT calculations based verification.

A major evaluation of our CubicGAN model is to check whether it can generate new cubic materials with novel prototypes, which are represented by distinct anonymized formulas in

Pymatgen. As shown in Table 3, we find that 24 and 1 novel prototypes for ternary and quaternary materials, respectively, have been found in our generated samples that are not existent in our training data. For relieving the burden of DFT calculations, we choose to remove the samples that contain lanthanoid and actinoid elements. In total, 1 064 650 ternary materials are left, of which 209 744 materials are of new crystal prototypes. The distribution of prototypes for 1 064 650 materials is shown in Figure S4, Supporting Information. Similarly, 4 382 130 quaternary materials are left after removing lanthanoid and actinoid elements, of which 260 891 materials are of the new crystal prototype (the prototype ABC_6D_6 –216). Since only two ABCD_6 –216 materials exist in the quaternary training dataset OQMD-QC1, we also include ABCD_6 –216 materials for the downstream DFT analysis considering the huge number of generated ABCD_6 –216 samples (1 655 407). After searching thoroughly in databases of OQMD, MP, and ICSD, only a limited number of materials with ABCD_6 –216 are found, as shown in Table S4, Supporting Information. Then, we perform charge neutrality check by Pymatgen and CGCNN formation energy filtering on 209 744 ternary materials and 1 916 298 ($260\,891 + 1\,655\,407$) quaternary materials. While each material might have different atom arrangements in the unit cell that maps to the same space group, in this work, we only choose one of them for DFT calculations. Finally, 17 303 ternary materials and 91 594 quaternary materials are left for DFT optimization. In total, 36847 candidate materials have been relaxed successfully with 14 433 ternary and 22 414 quaternary samples.

3.3. Discovery of 506 New-Prototype Stable Materials Verified with DFT Calculations

After filtering down materials with novel prototypes, we perform DFT optimization on materials with CGCNN-predicted negative formation energy, and we use Γ points and mechanic constants to further scale down the successfully relaxed structures. Phonon dispersion is the eventual criterion to determine the stability of structures.

3.3.1. Gamma Points and Mechanic Constants Filtering

The vibrational frequencies at the Γ point together with the elastic constants of screened structures were obtained by calculating the Hessian matrix (matrix of the second derivatives of the energy with respect to the atomic positions),^[48] which can be done by setting IBRION=6 (NFREE= 4) in VASP run. For cubic structures, the mechanical stability of lattice structures is verified as $C_{11} > 0$, $C_{44} > 0$, $C_{11} > |C_{12}|$, $C_{11} + 2C_{12} > 0$, where C_{ij} are components of elastic constant matrix.^[49] After screening the materials with the mechanical criteria, we further narrow-down the materials by checking the vibrational frequencies at the Γ point. All materials with negative Γ point frequencies were discarded.

3.3.2. Phonon Dispersion Calculation

After the structures pass the mechanical stability criteria and all Γ point frequencies are positive, we further calculate the full

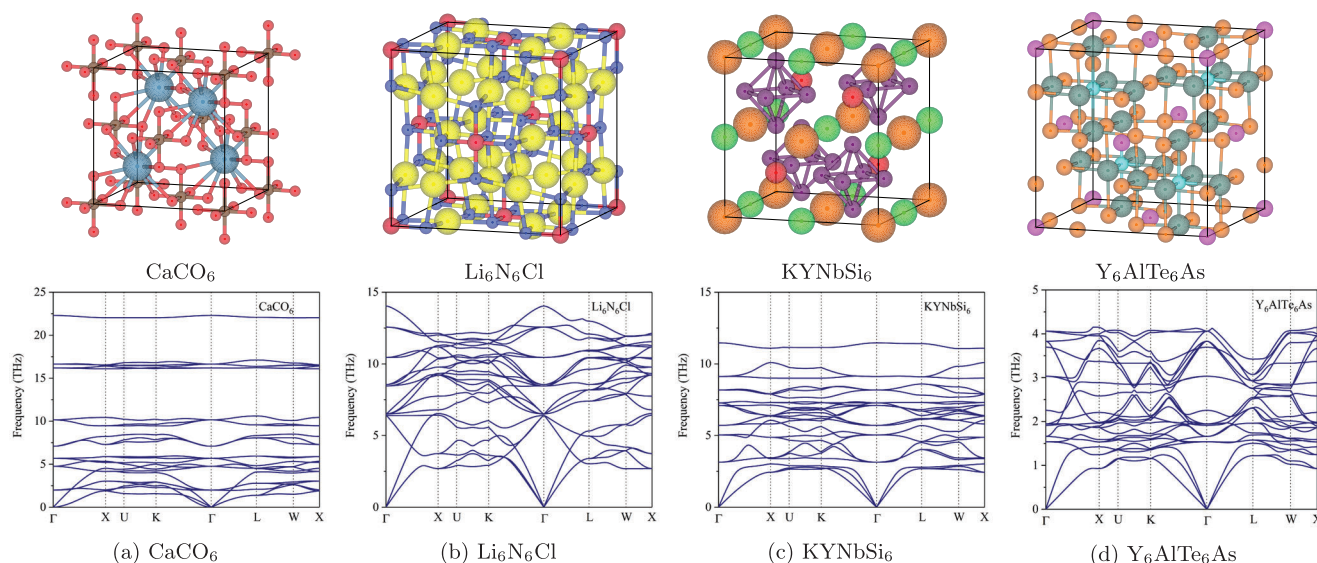


Figure 3. Examples of four new prototype materials. Top row is the crystal structures and the second row with corresponding phonon dispersion.

phonon dispersions in the first Brillouin zone (BZ). All 2nd interatomic force constants (IFCs) of the cubic structures were computed in a $2 \times 2 \times 2$ supercell based on their corresponding primitive cell. Then, the phonon dispersions were calculated by using the PHONOPY package^[50] with high symmetry paths $\Gamma \rightarrow X \rightarrow U \rightarrow K \rightarrow \Gamma \rightarrow L \rightarrow W \rightarrow X$.^[51]

In total, four prototypes with stable materials are discovered: ABC_6 –216, AB_6C_6 –225, $ABCD_6$ –216, and ABC_6D_6 –216. The details of the number of materials for each prototype are shown in Figure S5, Supporting Information. To our best of knowledge, ABC_6 –216 and ABC_6D_6 –216 are novel prototypes that are not in our training dataset, and the validation sets MP-TC3 and ICSD-TC3. Also, the AB_6C_6 –225 prototype is not in the training dataset and only one unstable material can be found in MP. However, our method finds 42 stable ones. Two materials of $ABCD_6$ –216 prototype are in the training dataset, and several others are in MP and ICSD. We expand the datasets by finding 62 stable materials of prototype $ABCD_6$ –216. Overall, we find 183 stable ternary materials and 323 stable quaternary materials. **Figure 3** shows four newly discovered stable cubic materials with their phonon dispersion curves. The CIF files of the 506 new prototype materials can be found in the Supporting Information.

Some interesting features have been observed from the phonon dispersions of newly discovered materials. For instance, a couple of hundred cubic structures we have screened out possess significant but tunable phonon bandgaps (e.g., $CaCO_6$ as shown in Figure 3a). Such phonon bandgaps could lead to extraordinary hot carrier performance,^[52–55] which is very promising for their potential application in photovoltaics, nonlinear optics (e.g., ultrashort pulsed lasers), multi-exciton generation devices, and even photocatalysis. Large phonon bandgaps at extremely high frequencies (such as H-containing materials not shown herein) deserve further investigation for their electron-phonon coupling properties,^[56–58] which could be beneficial for designing novel superconductors. Also, there are many cubic materials possessing very soft acoustic modes, for example, the longitudinal acoustic (LA) phonon branch in

$KYNbSi_6$ (Figure 3c), which indicate strong phonon anharmonicity and could be good candidates for waste-heat energy recovery (thermoelectrics). Last but not least, the phonon dispersion of Y_6AlTe_6As structure exhibits a very large gradient in high-frequency optical phonon modes and thus their phonon group velocities will be very high, which could lead to a significant contribution to the overall thermal transport from these optical modes and thus unusual temperature-dependent lattice thermal conductivity.^[59]

3.4. Visualization of the Relationship between New and Existing Prototypes within the Same Space group

To qualitatively evaluate how the new-prototype materials are structurally different from existing materials, we represent both sets of materials using simulated (for generated samples) and real X-ray diffraction (XRD) spectrum of dimension 901, which is a way to analyze the structures of inorganic materials. The simulated XRDs are calculated using the Pymatgen package.^[47] We then map all the materials' XRD vectors into a 2D space using the t-SNE algorithm,^[60] which are then plotted to visualize how existing and novel materials of the same space group 216 are distributed. **Figure 4** shows t-SNE embedding of existing materials in the training dataset and newly discovered ABC_6 –216 materials. From Figure 4a, we can find that new prototype materials (dark green dots) form distinct clusters, and there are apparent boundaries between known and unknown materials, which indicates that our model can generate materials beyond the scope of existing prototypes with significant structure deviations. Figure 4b shows a zoomed region of clusters of novel ABC_6 –216 materials, which implies that even samples of the same prototype can form structurally different clusters. For the other three prototypes, the distribution of known structures and our new-prototype structures are shown in Figures S6–S9, Supporting Information. In Figure S6, Supporting Information, we find that the new-prototype (AB_6C_6) materials are mostly located at the peripheral regions of known materials clusters, indicating

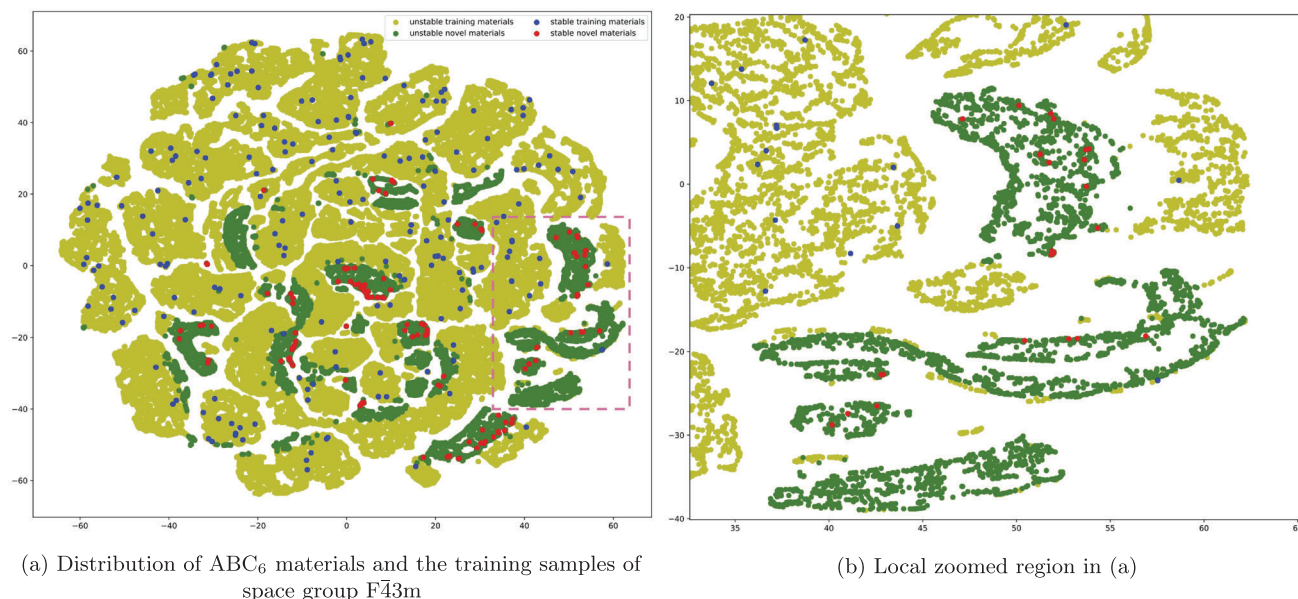


Figure 4. Visualization of the structural distributions of the materials in training set and the generated new-prototype ABC_6 materials both belonging to the space group of $F\bar{4}3m$. The new materials form structurally distinct clusters. The 2D space is generated by t-SNE embedding from high-dimensional XRD representations of cubic structures. a) Overall distributions of existing and new-prototype materials. b) Zoomed region as marked in sub-figure (a).

their structural closeness to known structures. In contrast, Figure S7, Supporting Information, shows that materials of two new-prototypes (ABC_6 and AB_6C_6) tend to form distinct clusters from known cubic materials in MP-TC3 and ICSD-TC3 validation sets, indicating their structural deviation from known materials. Additionally, for most of these new-prototype clusters, we have identified one or more DFT-verified stable materials. Figure S8, Supporting Information, shows that materials of new-prototype $ABCD_6$ form multiple new clusters, each of which contains multiple DFT-verified stable materials. Instead, materials of new-prototype ABC_6D_6 form much fewer cluster compared to the training set OQMD-TC3.

4. Conclusion

Large scale generation of new materials with distinct structures and functions are highly desirable for widely used high-throughput screening based materials discovery. Faced with astronomically large structural design space (compared to the space of the chemical compositions), the generator models have to exploit the implicit sophisticated physicochemical and geometric rules and constraints embedded in the existing crystal materials. Here we propose a novel GAN-based deep generative model for large-scale generation of three major types (space groups: 216, 225, 221) of cubic materials structures. Trained with 375 749 ternary cubic crystal structures from OQMD, our CubicGAN model can rediscover most of the known cubic structures as curated over more than 100 years of history within 10 million samplings. Especially, further analysis shows that our GAN model can generate not only new materials of existing prototypes but also new-prototype materials with distinct structural novelty. In total, we have identified 24 new prototypes of cubic materials. With rigorous DFT-based relaxation and phonon dispersion cal-

culation, we have identified and verified 506 new-prototype cubic materials, which are shared via our public database. From them, we have already identified several crystal structures with exceptional properties to be exploited in future. Together, our CubicGAN has demonstrated a promising path to large-scale generation and discovery of new materials.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Research reported in this work was supported in part by NSF under grant 1940099. This material is based upon work supported by the National Science Foundation Harnessing the Data Revolution Big Idea under Grant No. 1905775. The views, perspective, and content do not necessarily represent the official views of the NSF.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, J.H.; Methodology, Y.Z., J.H., and M.H.; Software, Y.Z., J.H., and Y.S.; Validation, Y.Z., M.H., J.H., D.S., and M.A.; Investigation, J.H., Y.Z., M.H., D.S., and Y.S.; Resources, J.H.; Data curation, Y.Z. and J.H.; Writing—original draft preparation, Y.Z., J.H., and M.H.; Writing—review and editing, J.H., M.H., Y.S., and A.N.; Visualization, Y.Z. and Y.S.; Supervision, J.H. and M.H.; Funding acquisition, J.H. and M.H.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

crystal structure generation, cubic crystals, deep neural networks, generative adversarial networks

Received: February 9, 2021

Revised: May 3, 2021

Published online: August 5, 2021

- [1] P. Wollmann, M. Leistner, U. Stoeck, R. Gr nker, K. Gedrich, N. Klein, O. Throl, W. Gr hlert, I. Senkovska, F. Dreisbach, S. Kaskel, *Chem. Commun.* **2011**, 47, 5151.
- [2] A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, E. J. Reed, *Energy Environ. Sci.* **2017**, 10, 306.
- [3] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, C. Wolverton, *Phys. Rev. Mater.* **2018**, 2, 123801.
- [4] M. C. Sorkun, S. Astruc, J. V. A. Koelman, S. Er, *npj Comput. Mater.* **2020**, 6, 106.
- [5] Y. Liu, S. Wang, A. M. Nolan, C. Ling, Y. Mo, *Adv. Energy Mater.* **2020**, 10, 2002356.
- [6] A. M. Nolan, Y. Zhu, X. He, Q. Bai, Y. Mo, *Joule* **2018**, 2, 2016.
- [7] A. Belsky, H. Mariette, L. K. Vicky, L. Peter, *Acta Crystallographica Section B: Structural Science* **2002**, 58, 364.
- [8] A. R. Oganov, L. Andriy, V. Mario, F. Gilles, in *CECAM-Workshop Lausanne* **2012**, pp. 22–26.
- [9] A. R. Oganov, C. J. Pickard, Q. Zhu, R. J. Needs, *Nat. Rev. Mater.* **2019**, 4, 331.
- [10] Y. Wang, J. Lv, L. Zhu, Y. Ma, *Phys. Rev. B* **2010**, 82, 094116.
- [11] C. W. Glass, A. R. Oganov, N. Hansen, *Comput. Phys. Commun.* **2006**, 175, 713.
- [12] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, J. Hu, *npj Comput. Mater.* **2020**, 6, 84.
- [13] T. Long, N. M. Fortunato, I. Opahle, Y. Zhang, I. Samathrakakis, C. Shen, O. Gutfleisch, H. Zhang, *arXiv:2007.11228* **2020**.
- [14] Z. Ren, J. Noh, S. Tian, F. Oviedo, G. Xing, Q. Liang, A. Aberle, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung, T. Buonassisi, *arXiv:2005.07609* **2020**.
- [15] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, *Matter* **2019**, 1, 1370.
- [16] S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik, Y. Jung, *arXiv:2004.01396* **2020**.
- [17] G. Hautier, C. Fischer, V. Ehlacher, A. Jain, G. Ceder, *Inorg. Chem.* **2011**, 50, 656.
- [18] J.-X. Shen, M. Horton, K. A. Persson, *npj Comput. Mater.* **2020**, 6, 161.
- [19] Y. Song, E. M. D. Siriwardane, Y. Zhao, J. Hu, *arXiv:2012.09314* **2020**.
- [20] J. Noh, G. H. Gu, S. Kim, Y. Jung, *Chem. Sci.* **2020**, 11, 4871.
- [21] Y. Jung, in *Telluride Workshop Machine Learning and Informatics for Chemistry and Materials*, Telluride Science Research Center, Telluride **2018**.
- [22] L. Wille, *Nature* **1986**, 324, 46.
- [23] D. J. Wales, J. P. Doye, *J. Phys. Chem. A* **1997**, 101, 5111.
- [24] S. Goedecker, *J. Chem. Phys.* **2004**, 120, 9911.
- [25] A. Noura, N. Sokolovska, J.-C. Crivello, *arXiv:1810.11203*, **2018**.
- [26] C. Court, B. Yildirim, A. Jain, J. M. Cole, *J. Chem. Inf. Model.* **2020**, 60, 4518.
- [27] V. Korolev, A. Mitrofanov, A. Eliseev, V. Tkachenko, *Mater. Horiz.* **2020**, 7, 2710.
- [28] G. E. Hinton, R. R. Salakhutdinov, *Science* **2006**, 313, 504.
- [29] D. P. Kingma, M. Welling, *arXiv:1312.6114* **2013**.
- [30] C. Doersch, *arXiv:1606.05908* **2016**.
- [31] M. Arjovsky, S. Chintala, L. Bottou, *arXiv:1701.07875*, **2017**.
- [32] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, in *Proc. of the IEEE Int. Conf. on Computer Vision*, IEEE, Piscataway, NJ **2017**, pp. 2223–2232.
- [33] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. R hl, C. Wolverton, *npj Comput. Mater.* **2015**, 1, 15010.
- [34] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *Appl. Mater.* **2013**, 1, 011002.
- [35] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, 65, 1501.
- [36] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, 120, 145301.
- [37] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, 31, 3564.
- [38] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, S.-C. Zhang, *Proc. Natl. Acad. Sci. U. S. A.* **2018**, 115, E6411.
- [39] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, 571, 95.
- [40] M. Mirza, S. Osindero, *arXiv:1411.1784* **2014**.
- [41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Adv. Neural Inf. Process. Syst.* **2014**, 2672.
- [42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, *Adv. Neural Inf. Process. Syst.* **2017**, 5767.
- [43] Y. Li, W. Yang, R. Dong, J. Hu, *arXiv:2010.16099* **2020**.
- [44] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, USENIX Association, Berkeley, CA **2016**, pp. 265–283.
- [45] F. Chollet, et al., Keras, <https://keras.io> **2015**. Last accessed on October, 2020.
- [46] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, 361, 360.
- [47] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, 68, 314.
- [48] Y. Le Page, P. Saxe, *Phys. Rev. B* **2002**, 65, 104104.
- [49] S. Zhang, R. Zhang, *Comput. Phys. Commun.* **2017**, 220, 403.
- [50] A. Togo, I. Tanaka, *Scr. Mater.* **2015**, 108, 1.
- [51] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, I. Tanaka, *Comput. Mater. Sci.* **2017**, 128, 140.
- [52] S. Chung, S. Shrestha, X. Wen, Y. Feng, N. Gupta, H. Xia, P. Yu, J. Tang, G. Conibeer, in *IOP Conf. Series: Materials Science and Engineering*, vol. 68, IOP Publishing, Bristol, England **2014**, 012002.
- [53] B. Thapa, M. Dubajic, M. P. Nielsen, R. Patterson, G. Conibeer, S. Shrestha, in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, Piscataway, NJ **2020** pp. 0793–0797.
- [54] X. Zhang, H.-Y. Song, X. C. Nie, S.-B. Liu, Y. Wang, C.-Y. Jiang, S.-Z. Zhao, G. Chen, J.-Q. Meng, Y.-X. Duan, H. Y. Liu, *Phys. Rev. B* **2019**, 99, 125141.
- [55] A. D. Wright, C. Verdi, R. L. Milot, G. E. Eperon, M. A. P rez-Osorio, H. J. Snaith, F. Giustino, M. B. Johnston, L. M. Herz, *Nat. Commun.* **2016**, 7, 11755.
- [56] S.-Y. Yue, L. Cheng, B. Liao, M. Hu, *Phys. Chem. Chem. Phys.* **2018**, 20, 27125.
- [57] J.-Y. Yang, M. Hu, *Phys. Chem. Chem. Phys.* **2018**, 20, 24222.
- [58] J.-Y. Yang, W. Zhang, C. Xu, J. Liu, L. Liu, M. Hu, *Int. J. Heat Mass Transfer* **2020**, 152, 119481.
- [59] G. Qin, Z. Qin, H. Wang, M. Hu, *Phys. Rev. B* **2017**, 95, 195416.
- [60] L. v. d. Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, 9, 2579.