Su *et al. J Mater Inf* 2024;4:20
DOI: 10.20517/jmi.2024.24

Journal of Materials Informatics

Research Article

Open Access

CGWGAN: crystal generative framework based on Wyckoff generative adversarial network

Tianhao Su[1,#], Bin Cao[2,#], Shunbo Hu[1], Musen Li[1], Tong-Yi Zhang[1,2,*]

[1]Materials Genome Institute, Shanghai University, Shanghai 200444, China.
[2]Guangzhou Municipal Key Laboratory of Materials Informatics, Advanced Materials Thrust, Sustainable Energy and Environment Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, Guangdong, China.
[#]Authors contributed equally.

[*]**Correspondence to:** Prof. Tong-Yi Zhang, Materials Genome Institute, Shanghai University, 333 Nanchen Road, Shanghai 200444, China. E-mail: zhangty@shu.edu.cn

How to cite this article: Su T, Cao B, Hu S, Li M, Zhang TY. CGWGAN: crystal generative framework based on Wyckoff generative adversarial network. *J Mater Inf* 2024;4:20. https://dx.doi.org/10.20517/jmi.2024.24

Received: 1 Jul 2024 **First Decision:** 30 Aug 2024 **Revised:** 26 Oct 2024 **Accepted:** 31 Oct 2024 **Published:** 7 Nov 2024

Academic Editors: Xingjun Liu, Miao Liu **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Discovering novel crystals is a highly effective way to develop new materials, though it presents significant challenges. Recently, many artificial intelligence (AI) generative methods have been developed to generate new crystals. In this work, we present a crystal generative framework based on Wyckoff generative adversarial network (CGWGAN) to efficiently discover novel crystals. The CGWGAN includes three modules: a generator of crystal templates, an atom-infill module, and a crystal screening module. The generator uses a generative adversarial network (GAN) to produce crystal templates embedded with asymmetry units (ASUs), space groups, lattice vectors, and the total number of atoms within the lattice cell, ensuring that the generated templates precisely match all requirements of crystals. These templates become crystal candidates after filling in atoms of different chemical elements. These candidates are screened by M3GNet and the passed ones are subjected to density functional theory (DFT)-based calculations to finally verify their stability. As a showcase, the CGWGAN successfully discovers seven novel crystals within the Ba-Ru-O system, demonstrating its effectiveness. This work provides a knowledge-guided Artificial Intelligence generative framework for accelerating crystal discovery.

Keywords: Crystal discovery, crystal symmetries, generative adversarial, space symmetry

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

INTRODUCTION

Crystals play extremely important roles in various scientific disciplines, including physics, chemistry, biology, *etc.*, especially in materials science and engineering. They embody distinct and precise structures of the atomic arrangement. A perfect crystal is characterized by a repeating basic unit, known as the crystal lattice, which, when translated in parallel, completely fills the whole three-dimensional (3D) space without any gaps. Crystals are categorized into 14 Bravais lattices, belonging to seven crystal systems, according to point group theory[1] and are further divided into 230 space groups[2]. In addition to crystal structure, chemical composition is also essential information of crystals. For a given molecule with a specific chemical formula, its atoms can arrange into entirely different crystal structures, a phenomenon known as isomerism. To fully define a crystal, one must know its structure, the types and arrangement of atoms, and the number of atoms, which leads to an extremely large variety of possible crystals. The arrangement of atoms within the crystal lattice is usually expressed by fractional coordinates and satisfies the crystal symmetry operations. In 1922, Wyckoff introduced the concept of Wyckoff positions, which summarize the equivalent positions within space groups. A Wyckoff position in a space group includes all sites whose site-symmetry groups are conjugate subgroups. Each Wyckoff position in a space group is labeled by a letter, known as the Wyckoff letter. The number of distinct Wyckoff positions in a space group is finite, with a maximum of 27 in space group *Pmmm*. In total, there are 1,731 Wyckoff positions across all space groups[3]. Based on Wyckoff positions, one can define an irreducible unit in a lattice cell, termed the asymmetry unit (ASU)[3]. An asymmetric unit of a space group is the smallest, indivisible part from which the entire lattice cell can be constructed by applying symmetry operations to it. Constructing a lattice cell from an ASU necessitates the precise determination of Wyckoff positions. However, in crystal generative deep learning models, atom sites are often generated with noise, making it challenging to achieve exact matches between atom sites across subgroups. The present work studies this issue deeply.

In addition to the commendable efforts in high-throughput computations and experiments[4], the data-driven exploration of novel crystals is rapidly advancing in tandem with the significant progress in artificial intelligence (AI) and machine learning (ML)[5]. Various ML models have been proposed to design and discover new crystals, demonstrating remarkable success. For example, the Universal Structure Predictor, Evolutionary Xtallography (USPEX), is an evolutionary algorithm-based method[6] and enables crystal structure prediction under arbitrary conditions of pressure and temperature, relying solely on the chemical composition of the material. Wang *et al.* developed a method for crystal structure prediction "from scratch" using the particle swarm optimization (PSO) algorithm[7]. This method efficiently minimizes free-energy surfaces, combining total-energy calculations with the PSO technique, and requires only chemical compositions to predict stable or metastable structures under specified external conditions, such as pressure. Shao *et al.* adopted a symmetry tree graph and an AI-based symmetry selection strategy to decompose the search space into Wyckoff position, thereby simplifying the prediction process[8]. Their method was validated by successful predictions in binary Lennard-Jones mixtures and high-pressure ice phases with over 100 atoms. Ryan *et al.* employed atomic fingerprint (AFP)[9], which represents purely topological relations between atoms in the crystal structure and reduces the 3D crystal structures to one-dimensional (1D) representations. They used a variational autoencoder (VAE) to discern the atoms of different elements based on the topology of their crystallographic environment[9]. The trained model is able to generate crystal templates in order to predict the likelihood of forming new compounds by placing elements into these structural templates in a combinatorial fashion. Zhao *et al.* developed the physics guided crystal generative model (PGCGM) for efficient crystal material design, incorporating physical constraints related to interatomic distances[10]. They conducted density functional theory (DFT) calculations to validate the generated structures in terms of thermodynamic stability. Xie *et al.* proposed the crystal diffusion variational autoencoder (CDVAE) model[11]. CDVAE incorporates SE(3)-equivariance[12] to learn from the

data distribution of stable crystals. During the diffusion process, it adjusts atomic positions toward lower energy states and updates atom types to satisfy bonding preferences between neighboring atoms. Zhu *et al.* proposed a framework that combines Wyckoff position encoding with DFT to generate symmetry-compliant, stable inorganic crystals, facilitating the discovery of novel materials and the accurate prediction of ground-state and polymorphic structures[13].

In AI crystal generation, a lattice cell is commonly characterized by a triplet of matrixes (or vectors), containing two types of fundamental components. The first type of components is topology- and position-related, such as the three lattice constants, fractional coordinates, and space groups. The second type of component concerns the arrangement of various types of atoms, a critical factor influencing both the energy of crystal formation and atomic bonding. The two steps can be integrated into a single model using either a conditional generative framework or sequence-based methods. In a conditional generative model, a well-pretrained model can capture the interaction between symmetry and energy. However, developing such a state-of-the-art (SOTA) model is quite challenging. The sequence strategy helps alleviate the model's training burden. Our study adopts a similar approach: we generate a crystal site topology that precisely adheres to the space group symmetry, referred to as a crystal template. Subsequently, atoms are positioned based on the potential energy surface.

Feedback from the potential energy surface is crucial during crystal generation. Relying solely on DFT calculations is highly resource-intensive and time-consuming. Therefore, the crystal generation framework also incorporates SOTA surrogate models for energy calculations. Chen *et al.* developed the universal MatErials graph network (MEGNet) for accurate property prediction of molecules and crystals, achieving an accuracy of 86.2% in crystal formation enthalpy prediction[14]. The universal graph deep learning interatomic potential (IAP), named M3GNet[15], uses graph neural networks (GNNs) with three-body interactions. After being trained on the comprehensive DFT crystal dataset from the Materials Project (MP), M3GNet has demonstrated its efficiency in predicting stable materials by identifying 1.8 million potentially stable materials from a screening of 31 million hypothetical crystal structures. Merchant *et al.* explored scaling deep learning for materials discovery and developed graph networks for materials exploration (GNoME)[16]. GNoME focused on large-scale active learning and structural prediction, achieving a prediction accuracy with a mean absolute error of 11 meV/atom for crystal formation energy and a stable prediction rate above 80%. This led to the discovery of over 2.2 million stable crystal structures. Furthermore, the application of transformers and attention-based algorithms has shown great promise in the study of crystals. For example, CrysalGPT[17] demonstrates exceptional sequence-to-sequence transferability while maintaining accurate time-series predictions. Additionally, the molecular set transformer[18], a computational screening tool, efficiently prioritizes molecular pairs capable of forming stable multicomponent structures, such as co-crystals.

This work develops the crystal generative framework based on Wyckoff generative adversarial network (CGWGAN). Figure 1 illustrates the architecture of CGWGAN, which consists of three modules: crystal template generation, atom filling, and energy/phonon calculations. In the crystal template generation module, a triplet of the asymmetric unit ($\mathcal{A}$), space group ($\mathcal{G}$), and lattice constant vector ($\mathcal{L}$) ensures the structural identity of the crystals. The ASU, with its Wyckoff positions, seamlessly integrates with the group theory-guaranteed symmetry and periodic translation conditions, establishing a novel Wyckoff self-fulfillment crystal generation framework. After training, the generator can produce crystal templates to enhance diversity in crystal generation. These templates are then filled with atoms of different and/or the same elements to create atoms-loaded crystals. The formation energy and phonon stability of these crystals are screened using M3GNet. Finally, the surviving crystals are evaluated through DFT and phonon

Figure 1. The architecture of the CGWGAN, including three modules of the crystal template generation, the atoms-filling in of different or the same atomic elements, and the DFT and phonon calculations. In the crystal template generation, a novel triplet of ASU ($\mathcal{A}$), space group ($\mathcal{G}$), and lattice constant vector ($\mathcal{L}$) ensure the structural identity of crystals. CGWGAN: Crystal generative framework based on Wyckoff generative adversarial network; DFT: density functional theory; ASU: asymmetry unit.

calculations to confirm their synthesis feasibility.

MATERIALS AND METHODS

A crystal can be decomposed into two key components: (1) the arrangement of sites, which reflects the lattice geometry constrained by the symmetry group; and (2) the atomic attributes, which determine the energy distribution and bonding between atoms. The generation of a CGWGAN model follows this sequence, beginning with the template (site) generation, followed by the atom infill steps. A crystal can be described by three fundamental elements: the asymmetric unit $\boldsymbol{\mathcal{A}}$; the space group $\boldsymbol{\mathcal{G}}$; and the lattice parameters $\boldsymbol{\mathcal{L}}$. These elements enable the derivation of all symmetry operations - including translations, rotations, reflections, and inversions - from any element $g \in \boldsymbol{\mathcal{G}}$. Using a finite set of group operations, a perfect crystal can be constructed in 3D space.

For a crystal sample $\boldsymbol{\mathcal{C}}^i$, the local site $encoder_L$ configuration is embedded as:

$$encoder_L(\boldsymbol{\mathcal{C}}^i) = \bigcup_{g \in \boldsymbol{\mathcal{G}}} g(\boldsymbol{r}_{\boldsymbol{\mathcal{A}}}) . \tag{1}$$

Here, $\bigcup_{g \in \boldsymbol{\mathcal{G}}} g(\boldsymbol{r}_{\boldsymbol{\mathcal{A}}})$ represents all the sites in the $\boldsymbol{\mathcal{A}}$ and g is a given space group from $\boldsymbol{\mathcal{G}}$.

$$\bigcup_{g \in \boldsymbol{\mathcal{G}}} g(\boldsymbol{r}_{\boldsymbol{\mathcal{A}}}) = Key_1\{\mathcal{H}, x, y, z\}^i \oplus Key_2\{\mathcal{H}, x, y, z\}^i \oplus \cdots, \tag{2}$$

where $\mathcal{H}$ represents the letter in Wyckoff set for a given space group, and x, y, z represent the corresponding fractional coordinates. Key_j signifies the embedding of ASU sites j = 1, 2, 3, …, k in the $\boldsymbol{\mathcal{A}}$, where k is a hyperparameter related to the number of total atoms in the crystal cell. In fact, once the specific space group

g and ASU sites are known, x, y, z can all be obtained through operation expansion and are not necessarily used as input when training the model. In addition to the site configuration, crystals are born with physical constraints on global site $encoder_G$:

$$encoder_G(\boldsymbol{C}^i) = g \oplus (E_1, E_2.., E_k) \oplus (a, b, c, \alpha, \beta, \gamma), \tag{3}$$

where E_j denotes atoms of different and/or same elements in $\boldsymbol{\mathcal{A}}$, $(a, b, c, \alpha, \beta, \gamma)$ represents lattice constant vector $\boldsymbol{\mathcal{L}}$. Therefore, a crystal can be fully represented as a vector $\widehat{C}$:

$$\widehat{\boldsymbol{C}}^i = encoder_L(\boldsymbol{C}^i) \oplus encoder_G(\boldsymbol{C}^i) \tag{4}$$

The atoms are masked in crystal templates,

$$\overline{\boldsymbol{C}}^i = Mask(\widehat{\boldsymbol{C}}^i) = \bigcup_{g \in \boldsymbol{\mathcal{G}}} g(\boldsymbol{r}_{\boldsymbol{\mathcal{A}}}) \oplus g \oplus (a, b, c, \alpha, \beta, \gamma). \tag{5}$$

The asymmetric unit $\boldsymbol{\mathcal{A}}$ is the minimal unit to represent a crystal geometry and atom sites. CGWGAN generates a crystal template based on the asymmetric unit $\boldsymbol{\mathcal{A}}$ with Wyckoff coordinates, space groups $\boldsymbol{\mathcal{G}}$ and the lattice vectors $\boldsymbol{\mathcal{L}}$.

Generative adversarial networks (GANs)[19,20] are founded on the principles of game theory and implemented using neural networks. The architecture of GAN is displaced in Figure 2, which has two components: generator and discriminator. The generator begins by sampling data from a latent space and, through a network, transforms the sampled data back into the original feature space, thereby generating fake data. Both the fake data and the real data are then fed into the discriminator. The discriminator's task is to distinguish between these two types of data. When the discriminator fails to differentiate fake data from real data, the GAN generates potential new data. CGWGAN uses Wasserstein Generative Adversarial Network[21] (WGAN) with a gradient penalty (GP) training strategy. CGWGAN loss $\mathcal{L}_g$ focuses solely on the overall coherence of the crystal templates,

$$\mathcal{L}_g = \mathbb{E}_{\boldsymbol{c} \sim p(\boldsymbol{c})}[f_w(\boldsymbol{c})] - \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_w\left(G_\theta(\boldsymbol{z})\right)\right] \tag{6}$$

where $\boldsymbol{c}$ is the real sample with distribution $p(c)$, z is a random noise with the $p(z)$ distribution, and G_θ and f_w represent the generator and discriminator, respectively. A self-attention module is integrated into the discriminator. Figure 2A illustrates the architecture of the generator, which consists of $N = 3$ fully connected layers with ReLU activation functions, followed by an output layer comprising a fully connected layer with a Tanh activation function. Figure 2B illustrates the configuration of the discriminator. The input crystal geometry undergoes, three times, fully connected layers with LeakyReLu and is subsequently fed into two sub-discriminators. The first sub-discriminator employs three 1D convolutional layers to transform the input vector into three distinct output vectors: $\boldsymbol{K}$ (Key), $\boldsymbol{Q}$ (Query), and $\boldsymbol{V}$ (Value)[22], thereby facilitating the computation of self-attention mechanisms. The second sub-discriminator is constructed similarly to a Residual Network, which serves to mitigate the issue of vanishing gradients. The information from the two sub-discriminators is aggregated by a fully connected layer, culminating in the final decision-making process.

All atom-filled crystals are screened first with those in the database (MP) to ensure that they are new crystals. Then, the pre-trained M3GNet model is adopted to examine their formation energy and kinetic stability, quickly. The crystal candidates that pass M3GNet are labeled with the formation energy and the

Figure 2. (A) Generator architecture configuration; (B) A 1D convolutional self-attention block with residual connections in the discriminator. 1D: One-dimensional.

lowest phonon frequency and stored in an ASE database. The corresponding automation scripts are provided in the data available section. After that, the Vienna ab initio simulation package (VASP), which employs DFT is employed to calculate the formation energy, with the exchange-correlation functional[23] of generalized gradient approximation (GGA). Detailed specifications of the cutoff energy, K-spacing, and convergence criteria for first-principles calculations are provided in the Supplementary Materials. The formation energy E_{form} is defined by

$$E_{\text{form}} = E_{\text{total}} - \sum n_i E_i, \tag{7}$$

where E_{total} is the total energy of the crystal, n_i and E_i are, respectively, the number and the energy of the i-th atom in its standard state. The structural reliability assessment is conducted through the calculation of phonon spectra utilizing the density functional perturbation theory (DFPT) method, in conjunction with the phonopy library[24]:

$$\omega(q,s) = \sqrt{\frac{1}{M}\Phi(q,s)}, \tag{8}$$

where $\omega(q, s)$ is the phonon frequency at wave vector q and phonon mode s, M is the atomic mass, and $\Phi(q, s)$ is the force constant matrix. The absence or weakness of imaginary frequencies in the phonon spectrum indicates that there is no tendency for spontaneous alteration in the atomic arrangement within the crystal structure, signifying a state of structural stability. Structures that pass both formation energy and phonon tests of DFT are novel crystals with great potential to be experimentally synthesized.

RESULTS AND DISCUSSION

This study uses 154,714 crystal structures from the MP database, encompassing 228 space groups, excluding space groups No168 and No207, across the seven crystal systems. Among these, 36,368 structures, exhibiting three and four asymmetric unit sites and possessing symmetry beyond P1, underwent screening. The distributions of 154,714 crystal structures in the 228 space groups and seven crystal systems are presented in Supplementary Figures 1 and 2.

The CGWGAN generates complete structures by first creating site templates and then filling them with suitable elements, making the generator inherently efficient. During the generation process, templates with inter-site distances less than 1 Å are automatically filtered out. A minimum distance of 1 Å is set as the lower limit because the smallest diameter of H atom is 1.06 Å[25], slightly larger than 1 Å, thereby structures with inter-site distances below this threshold are highly unlikely to form viable crystals after atomic infill. This tolerance significantly reduces computational cost without compromising the quality of the generated structures. Notably, 1 Å is a hyperparameter that can be adjusted based on the research system. This restriction effectively removes abnormal high-density structure prototypes. The atomic number density (ρ_A), expressed as $\frac{N_{atoms}}{V_{cell}}$, serves as a common indicator reflecting the distribution pattern of atoms in real space[26,27]. For instance, the CGWGAN generator produces a total of 140,000 crystal templates, each with a maximum of four Wyckoff positions, allowing for the accommodation of up to four elemental types. Supplementary Figure 3 categorizes these 140,000 crystal templates across all crystal systems. The distribution of synthetic crystal templates as a function of atomic number density (ρ_A), as shown in Figure 3A, overlaps with that of the training dataset by more than 93%, demonstrating a high consistency between the templates and the inherent distribution of the MP database. It is noteworthy that ρ_A was not explicitly provided to CGWGAN during the encoding and training process; however, CGWGAN successfully captured the underlying distribution pattern.

The alignment of features between synthetic templates and real crystals can be examined by analyzing the feature distributions in the latent space of the discriminator in CGWGAN. Figure 3B presents a t-distributed stochastic neighbor embedding (t-SNE) plot of the latent space features, demonstrating that the manifold features of the synthetic template structures closely follow those of the MP real crystals. All generated crystal templates are available as open-source on the Hugging Face platform.

The convergence percentage of generated crystals is used as a metric to evaluate our crystal generators. We use the popular physical method PyXtal[26] as a baseline. PyXtal automatically finds suitable combinations of Wyckoff positions using only the chemical composition and symmetry group information, employing a stepwise merging scheme. Using the open PyXtal package, we generated crystals ranging from just two atoms (Ni1Se1) to 32 atoms (P32) and, correspondingly, generated these crystals with CGWGAN. Figure 4A shows the convergence percentages of crystals, generated by CGWGAN and PyXtal, against the number of atoms in a unit cell. The convergence percentage of crystals generated by PyXtal decreases approximately and linearly with the increase in the number of atoms in a unit cell, as shown by the dashed blue line. The convergence percentage of crystals generated by CGWGAN is almost independent of the number of atoms in a unit cell. In ternary systems with more than 20 atoms per unit cell (e.g., $Mg_4Si_4O_{12}$), the convergence percentage of crystals generated by PyXtal is about 50%, while the crystals generated by CGWGAN maintain a convergence percentage of around 95%. Thus, the performance of crystal formation is almost independent of the number of atoms and the type of elements. Furthermore, CGWGAN achieves convergence percentages of 88% and 74%, respectively, for binary and mono-elemental crystal generation. The results show that only if the number of atoms per unit cell is less than 10 for mono-elemental crystal generation, PyXtal is better than CGWGAN, whereas for most cases, CGWGAN performs better than PyXtal. A detailed illustration of the outcomes for CGWGAN and PyXtal is provided in the Supplementary Materials.

Atoms of different elements are infilled into the crystal templates under the supervision of M3GNet to evaluate the formation energy and phonon spectrum of the candidate crystals. Candidates that pass the M3GNet screening have negative formation energies and no imaginary phonon modes in the entire Brillouin zone. These candidates then undergo final validation with CGWGAN through DFT calculations

Figure 3. (A) The distribution of the synthetic database and the MP database; (B) The latent space distribution after the 1D convolutional self-attention block of the discriminator. MP: Materials Project; 1D: one-dimensional.

Figure 4. (A) The convergence percentage versus the number of atoms in a unit cell of crystals generated by CGWGAN and Pyxtal, covering unary, binary, and ternary systems; (B) Filtering funnel diagram for the selection of Ba-Ru-O crystal candidates; (C) The distribution of formation energy calculated by the DFT relaxation on the Ba-Ru-O 1,156 crystal candidates. CGWGAN: Crystal generative framework based on Wyckoff generative adversarial network; DFT: density functional theory.

for formation energy and dynamic stability. DFT relaxation calculations identify energetically favorable atomic positions by allowing lattice constants and all atoms to shift slightly from their original locations. The convergence percentage is defined as the proportion of structures that converge under the calculation conditions of an energy convergence criterion of 1×10^{-7} eV, a maximum of 300 relaxation steps, and a maximum shift magnitude of 0.1 Å. Before phonon calculations, it is crucial to perform a higher-precision optimization of the candidate structures. This includes refining the k-point mesh in reciprocal space with a spacing of 0.2 $Å^{-1}$ and applying Van der Waals force corrections. All electronic minimization algorithms update all bands for simultaneous orbital updates with plane-wave cutoff energy based on the pseudopotential. Notably, due to the absence of specific symmetry information in space group P1, we further calculated six eigenvalues related to mechanical stability based on elastic stability criteria[28].

Figure 4B summarizes the crystal generation process for the Ba-Ru-O system through a sequential selection approach within the CGWGAN workflow. Initially, 120,000 crystal templates, each with a maximum of four Wyckoff positions, were generated. After inserting Ba-Ru-O atoms into the ASU sites, 1,795,032 crystal candidates were generated, as shown in Figure 4B. Subsequently, the generated crystal candidates were deduplicated against the Ba-Ru-O chemical compositions recorded in the MP dataset using the crystal matching function of Pymatgen[29]. This process removed two crystals already present in the MP database, leaving 1,795,030 newly discovered candidates. These Ba-Ru-O crystal candidates were then screened by M3GNet and ranked according to their predicted formation energy. Due to the computational complexity of phonon calculations, only the top 5,000 candidates with the lowest formation energies were selected for phonon spectrum calculations. However, only 1,156 of these 5,000 candidates passed the M3GNet phonon spectrum stability criteria. The phonon spectrum stability screened by M3GNet is based on the criterion that the minimum phonon frequency should be greater than -0.5 THz. DFT relaxation was then performed on the 1,156 crystal candidates. Figure 4C shows the distribution of their formation energies, all of which are negative. Phonon frequencies were calculated using DFPT[30] by establishing a 2 × 2 supercell. The stability criterion in the DFT-based phonon spectrum is greater than 0 THz. Finally, the top 30 candidates with the lowest energies underwent rigorous phonon spectrum calculations, resulting in seven newly discovered crystals.

Table 1 lists the seven generated novel computational crystals. Using Pymatgen symmetry checking, we found that these novel structures correspond to four crystal systems and seven space groups: Triclinic (#1), Orthorhombic (#25, #35, #71), Tetragonal (#99, #123) and Trigonal (#166). Additionally, the structures were verified using AFLOW's XtalFinder, and the results are provided in the Supplementary Materials.

Figure 5 presents the phonon spectra for these candidate crystals, corresponding to space group numbers: (a) 1, (b) 25, (c) 35, (d) 71, (e) 99, (f) 123, (g) 166. For the crystal structures shown in Figure 5E and F, only minor discrepancies are observed in the phonon spectra. Additional X-ray diffraction (XRD) simulations using WPEM software[32], as depicted in Figure 5H, confirm that these structures correspond to different configurations with varying asymmetric unit compositions in distinct space groups. This demonstrates that CGWGAN algorithm effectively learns the deeper rules of crystal geometry distribution. Apart from the structures shown in Figure 5, numerous metastable structures were also generated and most of them have negative formation energies but exhibit imaginary frequencies at some high-symmetry points in phonon calculations.

Some of the newly discovered crystals might belong to the superoxides[33,34], whose valences might be higher than the typical oxidation states observed in common oxides. The Bader method[35,36] is widely used to partition charges to each ion. For example, the charges in the newly discovered crystal $BaRuO_6$ are

Table 1. The chemical formulas, space group, and formation energy of the structures were determined following high-precision optimization in the low-energy state

Compositions	Space groups	Formation energy
$BaRuO_6$	*P*1 (1)	-2.01 eV
$BaRuO_3$	*Pmm*2 (25)	-1.93 eV
$BaRu_2O_3$	*Cmm*2 (35)	-1.94 eV
$BaRuO_2$	*Immm* (71)	-1.87 eV
$BaRuO_3$	*P*4*mm* (99)	-1.97 eV
$BaRuO_3$	*P*4/*mmm* (123)	-1.97 eV
$Ba_3Ru_3O_9$	R$\bar{3}$*m* (166)	-2.04 eV

Figure 5. Stable phonon spectra of generated samples in Ba-Ru-O system: this calculation is conducted on a supercell two times along each axis. Visualization is conducted using VESTA[31].

partitioned by the Bader charge analysis. The valence electrons are 11 e for Ba, 13 e for Ru, and 6 e for O. Figure 6A visualizes the isosurface of charge density in the irreducible lattice cell of $BaRuO_6$, including one Ba atom, one Ru atom and six O atoms. Figure 6B shows the corresponding electronic band structure, indicating that the O atoms between O_2-O_6 and O_1-O_5 exhibit significant electron cloud overlap, whereas O_3 and O_4 do not significantly exchange electrons with other O atoms and acquire more effective charge. Furthermore, the bands in the electronic band structure near the Fermi level are mainly contributed by O atoms with a minor contribution from Ru atoms, while Ba remains at a deep level; consequently, no charge associated with Ba is visualized on the isosurface of charge density. Table 2 lists fractional charge to each ion inside the crystal from the Bader charge analysis.

Table 2. The valence charge of each ion in $BaRuO_6$ based on Bader charge analysis

Ion	Valence charge (|e|)	Volume (Å^3)
Ba	2.671630	22.353047
Ru	0.897347	10.919872
O_1	-0.465287	11.005901
O_2	-0.345221	10.729654
O_3	-0.893444	14.835121
O_4	-0.843930	15.366585
O_5	-0.505435	11.946478
O_6	-0.515659	12.959701

Figure 6. (A) The isosurface charge density in the irreducible lattice cell of $BaRuO_6$; and (B) the corresponding electronic band structure.

First-principles calculations select generated crystals with a negative formation energy, ensuring that the generated crystals are thermodynamically stable at zero K. Phonon spectrum calculations guarantee that the generated crystals are mechanically stable at zero K. These two properties are essential and necessary conditions for experimental synthesis of these generated crystals. In general, experimental synthesis of crystals is analyzed by two aspects. The first is the thermodynamic aspect that the Gibbs free energy of products (synthesized crystals) must be lower than the Gibbs free energy of reactants (raw materials). The second is the kinetic aspect that the resistance of paths from raw materials to products must be low, which is usually represented by energy barriers. Both synthesis thermodynamics and kinetics can be varied by adjusting the synthesis temperature, pressure, and other applied electric, magnetic, or microwave fields, and by utilizing catalysts. All of these are beyond the scope of the present study.

Regarding the dataset distribution used in this study, the generated distributions, statistical details on generated crystals and relaxed lattice parameters, a step-by-step breakdown of CGWGAN's generative process, comparisons with other models or algorithms, and information on computational efficiencies, please refer to the Supplementary Materials.

CONCLUSIONS

In summary, CGWGAN has demonstrated its capability to efficiently generate high-quality crystal structures of target crystal systems and space groups. As an example, the present work discovers seven novel crystals within the Ba-Ru-O system, boasting high thermodynamic and kinetic stability. Obtaining the ASU through known operations in group theory, this physical encoding enables a fully reversible generative process. Beyond the architecture applied in the paper based on generative adversarial neural networks, this framework is versatile and applicable to any generative model framework, including VAEs and diffusion models. Apart from crystal generation, CGWGAN can be easily integrated with AI XRD analyzer[37] service as out-library phase identification methods, thereby considerably fostering the development of materials informatics. The current CGWGAN has not paid sufficient attention to material properties, while the dream of material scientists and engineers is to design and discover novel crystals based on the need of material properties in scientific and engineering practice. To materialize this dream, CGWGAN will be improved by adding the conditions of certain designed material properties. In addition to M3GNet which is trained on the MP database, experimental crystal databases will be used to enhance the synthesizability of generated crystals from CGWGAN.

DECLARATIONS

Authors' contributions

Conceived the idea and designed the project: Su T, Cao B, Li M, Zhang TY
Performed data analysis and interpretation: Su T, Cao B, Hu S
Supervised the project, drafted the manuscript: Su T, Cao B, Zhang TY
Revised and finalized the manuscript: Su T, Cao B, Li M, Hu S, Zhang TY
All authors read and approved the final manuscript.

Availability of data and materials

The crystal templates generated are openly accessible at https://huggingface.co/datasets/caobin/CGWGAN. The program for the computation of energy and phonon spectra via M3GNet, an automated post-processing and repository storage system, is openly accessible at https://github.com/WPEM/CGWGAN.

Financial support and sponsorship

Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0003 and No. 2023A03J0103), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 12404276), the Special Funds of the National Natural Science Foundation of China (Grant No. 12347164), the China Postdoctoral, Science Foundation (Grant No. 2024T170541), China Postdoctoral Science Foundation (Grant No. GZC20231535), and National Key R&D Program of China (Grant No. 2023YFB4402600).

Conflicts of interest

Zhang TY is the Editor-in-Chief of the *Journal of Materials Informatics*, while the other authors have declared that they have no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2024.

REFERENCES

1. Togo A, Shinohara K, Tanaka I. Spglib: a software library for crystal symmetry search. arXiv. [Preprint.] Mar 13, 2024 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.1808.01590.
2. Pidcock E, Motherwell WDS, Cole JC. A database survey of molecular and crystallographic symmetry. *Acta Crystallogr B* 2003;59:634-40. DOI PubMed
3. Wondratschek H, Aroyo MI. 1.2 Crystallographic symmetry. In: International Tables for Crystallography. 2016. DOI
4. Cao B, Yang S, Sun A, Dong Z, Zhang TY. Domain knowledge-guided interpretive machine learning: formula discovery for the oxidation behavior of ferritic-martensitic steels in supercritical water. *J Mater Inf* 2022;2:4. DOI
5. Curtarolo S, Setyawan W, Wang S, et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput *ab initio* calculations. *Comput Mater Sci* 2012;58:227-35. DOI
6. Glass CW, Oganov AR, Hansen N. USPEX - Evolutionary crystal structure prediction. *Comput Phys Commun* 2006;175:713-20. DOI
7. Wang Y, Lv J, Zhu L, Ma Y. CALYPSO: a method for crystal structure prediction. *Comput Phys Commun* 2012;183:2063-70. DOI
8. Shao X, Lv J, Liu P, et al. A symmetry-orientated divide-and-conquer method for crystal structure prediction. *J Chem Phys* 2022;156:014105. DOI PubMed
9. Ryan K, Lengyel J, Shatruk M. Crystal structure prediction via deep learning. *J Am Chem Soc* 2018;140:10158-68. DOI PubMed
10. Zhao Y, Siriwardane EMD, Wu Z, et al. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Comput Mater* 2023;9:38. DOI
11. Xie T, Fu X, Ganea OE, Barzilay R, Jaakkola T. Crystal diffusion variational autoencoder for periodic material generation. arXiv. [Preprint.] Mar 14, 2022 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.2110.06197.
12. Batzner S, Musaelian A, Sun L, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* 2022;13:2453. DOI PubMed PMC
13. Zhu R, Nong W, Yamazaki S, Hippalgaonkar KH. WyCryst: Wyckoff inorganic crystal generator framework. *Mater* 2024;7:3469-88. DOI
14. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater* 2019;31:3564-72. DOI
15. Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci* 2022;2:718-28. DOI
16. Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature* 2023;624:80-5. DOI PubMed PMC
17. Sitapure N, Kwon JSI. CrystalGPT: enhancing system-to-system transferability in crystallization prediction and control using time-series-transformers. *Comput Chem Eng* 2023;177:108339. DOI
18. Vriza A, Sovago I, Widdowson D, Kurlin V, Wood PA, Dyer MS. Molecular set transformer: attending to the co-crystals in the Cambridge structural database. *Dig Discov* 2022;1:834-50. DOI
19. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv. [Preprint.] Dec 6, 2017 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.1701.07875.
20. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. arXiv. [Preprint.] Jun 10, 2014 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.1406.2661.
21. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. arXiv. [Preprint.] Dec 25, 2017 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.1704.00028.
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. [Preprint.] Aug 2, 2023 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.1706.03762.
23. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett* 1996;77:3865-8. DOI PubMed
24. Togo A. First-principles Phonon Calculations with Phonopy and Phono3py. *J Phys Soc Jpn* 2023;92:012001. DOI
25. Bohr N. On the constitution of atoms and molecules. 1913. Available from: https://www.nba-old.nbi.dk/pdffiles/trilogypart3.pdf. [Last accessed on 4 Nov 2024].
26. Fredericks S, Parrish K, Sayre D, Zhu Q. PyXtal: a Python library for crystal structure generation and symmetry analysis. *Comput Phys Commun* 2021;261:107810. DOI
27. Steed KM, Steed JW. Packing problems: high Z' crystal structures and their relationship to cocrystals, inclusion compounds, and polymorphism. *Chem Rev* 2015;115:2895-933. DOI PubMed
28. Mouhat F, Coudert FX. Necessary and sufficient elastic stability conditions in various crystal systems. *Phys Rev B* 2014;90:224104. DOI
29. Ong SP, Richards WD, Jain A, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput Mater Sci* 2013;68:314-9. DOI
30. Togo A, Tanaka I. First principles phonon calculations in materials science. *Scr Mater* 2015;108:1-5. DOI
31. Momma K, Izumi F. *VESTA*: a three-dimensional visualization system for electronic and structural analysis. *J Appl Crystallogr* 2008;41:653-8. DOI

32. Cao B. Whole pattern fitting of powder X-ray diffraction by expectation maximum algorithm. 2024. Available from: https://figshare.com/articles/code/Whole_Pattern_fitting_of_powder_X-ray_diffraction_by_Expectation_Maximum_algorithm/25060175?file=44225531. [Last accessed on 4 Nov 2024].
33. Hesse W, Jansen M, Schnick W. Recent results in solid state chemistry of ionic ozonides, hyperoxides, and peroxides. *Prog Solid State Ch* 1989;19:47-110. DOI
34. Hayyan M, Hashim MA, AlNashef IM. Superoxide ion: generation and chemical implications. *Chem Rev* 2016;116:3029-85. DOI
35. Henkelman G, Arnaldsson A, Jónsson H. A fast and robust algorithm for Bader decomposition of charge density. *Comput Mater Sci* 2006;36:354-60. DOI
36. Sanville E, Kenny SD, Smith R, Henkelman G. Improved grid-based algorithm for Bader charge allocation. *J Comput Chem* 2007;28:899-908. DOI
37. Cao B, Liu Y, Zheng Z, Tan R, Li J, Zhang T. SimXRD-4M: big simulated X-ray diffraction data accelerate the crystalline symmetry classification. arXiv. [Preprint.] Jun 15, 2024 [accessed on 2024 Nov 4]. Available from: https://doi.org/10.48550/arXiv.2406.15469.