

# VQCrystal: Leveraging Vector Quantization for Discovery of Stable Crystal Structures

ZiJie Qiu,<sup>1,\*</sup> Luozhijie Jin,<sup>2,\*</sup> Zijian Du,<sup>3</sup> Hongyu Chen,<sup>2,4</sup>

Yan Cen,<sup>3,†</sup> Siqi Sun,<sup>1</sup> Yongfeng Mei,<sup>5</sup> and Hao Zhang<sup>2,4,6,‡</sup>

<sup>1</sup>*Research Institute of Intelligent Complex Systems,  
Fudan University, Shanghai, China*

<sup>2</sup>*School of Information Science and Technology,  
Fudan University, Shanghai 200433, China*

<sup>3</sup>*Department of Physics, Fudan University, Shanghai 200433, China*

<sup>4</sup>*Department of Optical Science and Engineering and Key Laboratory  
of Micro and Nano Photonic Structures (Ministry of Education),  
Fudan University, Shanghai 200433, China*

<sup>5</sup>*Department of Materials, Fudan University, Shanghai 200433, China.*

<sup>6</sup>*State Key Laboratory of Photovoltaic Science and Technology,  
Fudan University, Shanghai 200433, China*

## Abstract

Discovering functional crystalline materials through computational methods remains a formidable challenge in materials science. Here, we introduce VQCrystal, an innovative deep learning framework that leverages discrete latent representations to overcome key limitations in current approaches to crystal generation and inverse design. VQCrystal employs a hierarchical VQ-VAE architecture to encode global and atom-level crystal features, coupled with an machine learning-based inter-atomic potential(IAP) model and a genetic algorithm to realize property-targeted inverse design. Benchmark evaluations on diverse datasets demonstrate VQCrystal’s advanced capabilities in representation learning and novel crystal discovery. Notably, VQCrystal achieves state-of-the-art performance with 91.93% force validity and a Fréchet Distance of 0.152 on MP-20, indicating both strong validity and high diversity in the sampling process. To demonstrate real-world applicability, we apply VQCrystal for both 3D and 2D material design. For 3D materials, the density-functional theory validation confirmed that 62.22% of bandgaps and 99% of formation energies of the 56 filtered materials matched the target range. Moreover, 437 generated materials were validated as existing entries in the full database outside the training set. For the discovery of 2D materials, 73.91% of 23 filtered structures exhibited high stability with formation energies below -1 eV/atom. Our results highlight VQCrystal’s potential to accelerate the discovery of novel materials with tailored properties.

## I. INTRODUCTION

The discovery of new functional materials through computational methods represents a frontier in materials science. Despite centuries of exploration, human has only scratched the surface of the vast material search space, with an estimated  $10^5 - 10^{61-3}$  order out of  $10^{104}$  theoretically possible solid inorganic materials having been identified to date. Expanding our catalogue of known materials is crucial for scientific advancement, particularly as data-driven research methodologies become increasingly integral to modern materials science. Advancements in first-principles calculations have accelerated crystal discovery, a prevalent framework combines high-throughput virtual screening (HTVS)<sup>5</sup> combined with density functional theory (DFT)<sup>6</sup>. This approach substitutes atoms in known structures, followed by DFT relaxation to assess stability, leading to databases like the Materials Project (MP)<sup>7</sup> and OQMD<sup>8</sup>. While accurate, DFT's computational demands limit large-scale applications, highlighting the need for more efficient methods.

Deep learning models offer a computationally efficient alternative to traditional first-principles calculations by generating new crystals through sampling from learned distributions. Methods<sup>9-12</sup> based on Generative adversarial network (GAN)<sup>13</sup> drives a similar sampling distribution to the database distribution with a generator and a discriminator. But the inherent training instability and low sampling diversity restrict their application to specific subsets of crystalline materials like those with space group 225<sup>10,11</sup>, binary Bi-Se systems<sup>9</sup>, or alloys<sup>12</sup>. More general methods of discovering crystals are based on variational autoencoder (VAE)<sup>14</sup> and diffusion model<sup>15</sup>. These models map the complex information of diverse unit cells onto a unified latent space, which enables the encoding and sampling of a wide variety of materials using a single model. Two notable examples of them are Fourier-transformed crystal properties framework (FTCP)<sup>16</sup> and Crystal Diffusion Variational AutoEncoder (CDVAE)<sup>17</sup>. FTCP uses a variational autoencoder model with invertible representation for crystal generation, incorporating composition and structure. The inverse design process utilizes a property-prediction head, with subsequent structure relaxation to enhance validity. But FCTP struggles with reconstruction and sampling validity. As an improvement, CDVAE uses a hybrid structure of VAE and diffusion models, using VAE for representation and a score-based diffusion model to iteratively refine structures, mimicking DFT relaxation. However, it still faces challenges in reconstruction and sampling validity, and lacks inverse design capability.

The development of deep learning pipelines for crystalline materials discovery and

inverse design faces three primary challenges: (1) Effective representation learning that facilitates bidirectional mapping between the crystal search space and a unified latent space. (2) The ability to perform approximate structure relaxation through neural networks, thereby enhancing sampling reliability. (3) Integration of a property prediction module and appropriate optimization algorithms for inverse design tasks. Current models have yet to successfully address these challenges simultaneously.

In this study, we introduce VQCystal, an innovative framework for the design of crystalline materials that addresses all the three primary challenges mentioned above. To the best of our knowledge, VQCystal is the first deep generative model that employs a hierarchical Vector Quantized Variational Autoencoder (VQ-VAE) architecture to encode the global and atom-level crystal features, which is an established technique for enhanced representation learning in image processing, molecular modeling, and point cloud analysis. This intuition also aligns with the discrete nature of crystal structures including finite symmetry operations, 255 distinct space groups<sup>18</sup>, and defined Wyckoff positions<sup>19</sup>. Additionally, VQCystal leverages OpenLAM<sup>20</sup>, an established machine learning toolkit, for structural relaxation decoupled from the tasks of representation learning. For inverse design, VQCystal is trained concurrently with an auxiliary task of predicting properties using the discretized global latent variable. During the sampling procedure, a Genetic Algorithm (GA) operating on codebook indices is employed to search for crystals with desired properties.

To benchmark the capabilities of VQCystal, three open benchmark datasets, MP-20<sup>21</sup>, Perov-5<sup>22</sup>, and Carbon-24<sup>23,24</sup> were tested. Compared its performance against state-of-the-art deep learning models for crystal generation, VQCystal achieved the highest validity and match rate, with 65.30% match rate, 100% structure validity, 84.58% composition validity, and 91.93% force validity on MP-20, with the best diversity with the Fréchet Distance (FD)<sup>25</sup> score of 0.152. Subsequent analyses show that the global and local latent space of VQCystal are highly interpretable. To demonstrate VQCystal’s applicability to real-world material design, two specific cases were explored: 3D crystalline materials and 2D crystalline materials. 56 out of the 20,789 3D crystals generated by VQCystal trained on the MP-20 database<sup>21</sup> were selected after removing duplicates, lanthanides and a neural-network-based<sup>26,27</sup> filtering under the criteria of bandgap ( $E_g$ ) between 0.5 and 2.5 eV and formation energy ( $E_f$ ) below -0.5 eV/atom. DFT validation showed that 62.22% of the bandgaps and 99% of the formation energies matched the target range.

Among the 20789 crystals, 437 materials, distinct from the training set, were validated by the dataset as duplicates of entries in the full database, with an average RMS distance of only 0.0509. For 2D materials, VQCystal was applied to the C2DB database<sup>28</sup>, generating 12,000 structures. After the similar filtering processes above, 73.91% of the 23 filtered relaxed materials had formation energies below -1 eV/atom, indicating high chemical stability.

## II. MODEL OVERVIEW

### A. VQCystal Model

The VQCystal shown in Figure 1 employs a hierarchical vector quantization architecture, which consists of three main components: the encoder, the vector quantization module, and the decoder, followed by auxiliary parts such as the property prediction head. The encoder in Figure 1(a) is composed of a hierarchical network that extracts both local and global information from the crystal. The crystal is represented by a tuple consisting of the atomic number of the L atoms, their respective frac-coordinates, and the unit cell basis vectors. The local feature  $\hat{z}_l$  is captured using a Transformer-based structure<sup>29</sup>, while the global feature  $\hat{z}_g$  is obtained by summing two components: One part is extracted by applying a SE(3)-equivariant periodic graph neural network (GNN)<sup>30</sup>, CSPNet shown in Figure 1(d), to the input crystal to extract unified features, and the other part is derived by applying a Graph Convolutional Networks (GCN)<sup>4</sup> to the local features to extract further information. These two components are summed together, followed by a pooling operation to get the final output  $\hat{z}_g$ . The use of SE(3)-equivariant graph networks for information extraction allows the model to effectively capture rotational and translational symmetries, making it ideal for handling crystalline structures.

The hierarchical Vector Quantization (VQ) module introduces discrete latent spaces and leverages a two-tiered approach, incorporating Residual Quantization (RQ)<sup>31</sup> techniques to efficiently compress the latent representations while preserving critical information. The VQ module handles both local and global features, quantizing them into discrete representation space as  $z_l$  and  $z_g$ . Stochastic sampling of codes, shared codebooks, and k-means clustering initialization enhance the performance and stability of the VQ module. Figure 1(b) shows the codebook space of local and global latents trained on the materials project datasets<sup>7</sup>. The decoder demonstrated in Figure 1(c) recon-

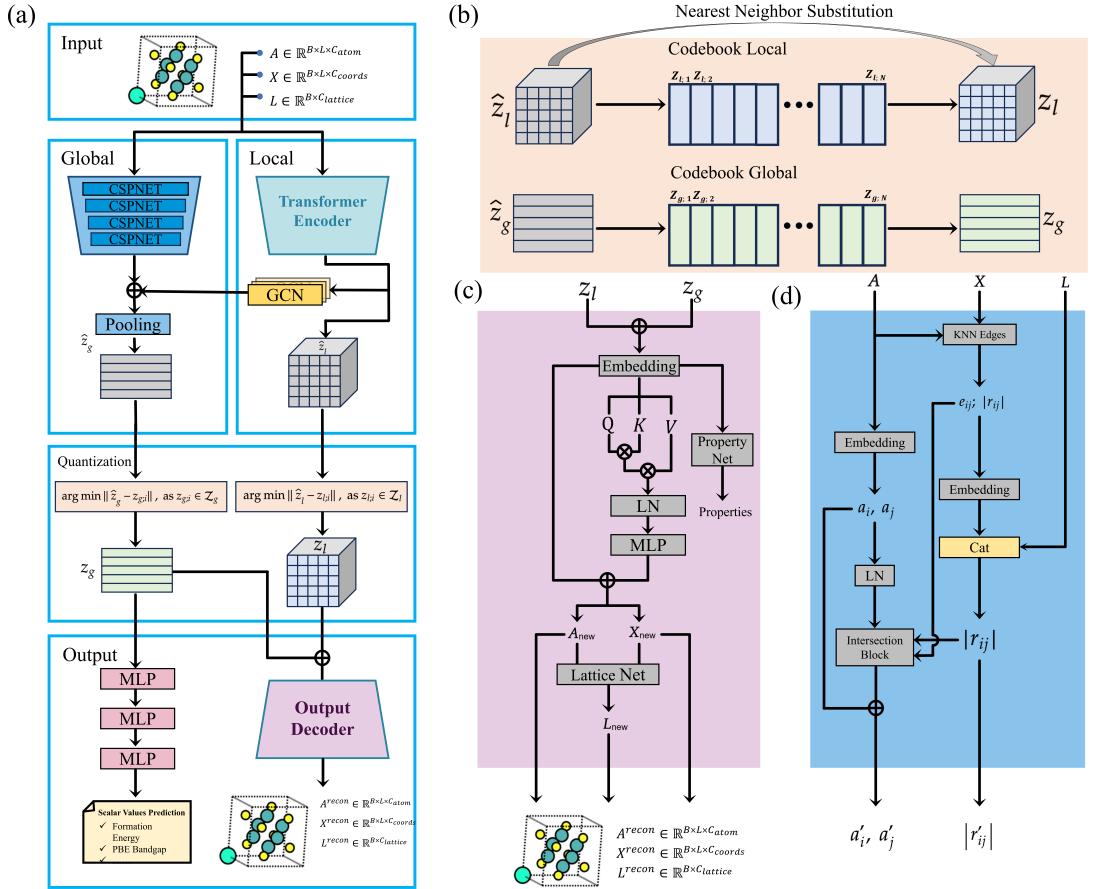


FIG. 1: (a) Overview of the VQCystal model. (b) Visualization of the local and global codebook space after PCA, trained on the Materials Project dataset. (c) The details of the decoder component. (d) The details of the CSPNet component.

structs the original input from the quantized latent representations  $z_l$  and  $z_g$ , using a transformer-based structure. The lattice parameters are predicted using a multilayer perceptron (MLP) after reconstructing the atoms and fractional coordinates. Additionally, the concatenation of  $z_l$  and  $z_g$  is passed through an MLP-based property network to predict various properties, such as formation energy, bandgap, and so on to ensure the latent contains several property information. The details of the VQCystal model is shown in supplementary information.

Despite incorporating Transformer layers, the overall complexity of VQCystal is dominated by the graph neural networks (GCN and CSPNet) with a time complexity of  $O(L \cdot (|E| \cdot d^2 + n \cdot d^2))$ , where  $L$  is the number of layers,  $|E|$  is the number of edges,  $n$  is the number of atoms, and  $d$  is the dimensionality of the features. For datasets like MP-20<sup>21</sup>, where  $n$  is much smaller than  $d$ , the effective complexity is dominated by terms involving  $d^2$  rather than the self-attention terms involving  $n^2$ , allowing VQCystal to efficiently handle sampling tasks. Full analysis of time complexity is shown in the supplementary information.

## B. Sampling Strategy

The VQCystal framework employs a sampling pipeline comprising two critical stages: (1) codebook indices search and (2) post-optimization. Within this framework, each crystal structure is uniquely represented by a pair of codebook indices,  $(I_{\text{global}}, I_{\text{local}})$ , corresponding to global and local structural features, respectively. The global index,  $I_{\text{global}}$ , is defined as an array in  $\mathbb{R}^{D_{\text{global}}}$ , while the local index,  $I_{\text{local}}$ , is characterized as an array in  $\mathbb{R}^{N \times D_{\text{local}}}$ . In this notation,  $D_{\text{global}}$  and  $D_{\text{local}}$  represent the number of global and local quantizers, respectively, and  $N$  denotes the maximum number of atoms in the crystal structure. The sampling process commences with the random selection of a crystal from the database, whereupon its  $I_{\text{local}}$  is fixed, leaving  $I_{\text{global}}$  as the sole variable for optimization. This strategic approach significantly constrains the search space, enhancing computational efficiency. Subsequently, a genetic algorithm is applied to optimize  $I_{\text{global}}$ , employing a suite of evolutionary operators including mutation, crossover, and selection. The objective of this optimization is to identify  $I_{\text{global}}$  values that, when decoded, yield crystal structures with minimal total energy, as estimated by the OpenLAM framework. Subsequently, a post-optimization phase is initiated. This stage utilizes OpenLAM to perform structural relaxation on the selected crystal candidates. The process culminates in the retention of only those structures that satisfy two stringent criteria: an estimated formation energy  $E_{\text{form}} < 0$  and a maximum atomic force  $f_{\text{max}} < 0.05 \text{ eV}/\text{\AA}$ .

In practical sampling, VQCystal significantly outperforms other models. Sampling a single crystal from the MP-20 dataset takes less than 0.1 seconds, and with the inclusion of structural relaxation and genetic algorithm updates through OpenLAM, the total time is around 15 seconds per sample. In contrast, USPEX<sup>32</sup> takes an average of 12.5 hours per sample, GN-OA<sup>33</sup> averages 3 minutes, and models like M3GNet<sup>34</sup>, DiffCSP<sup>30</sup>,

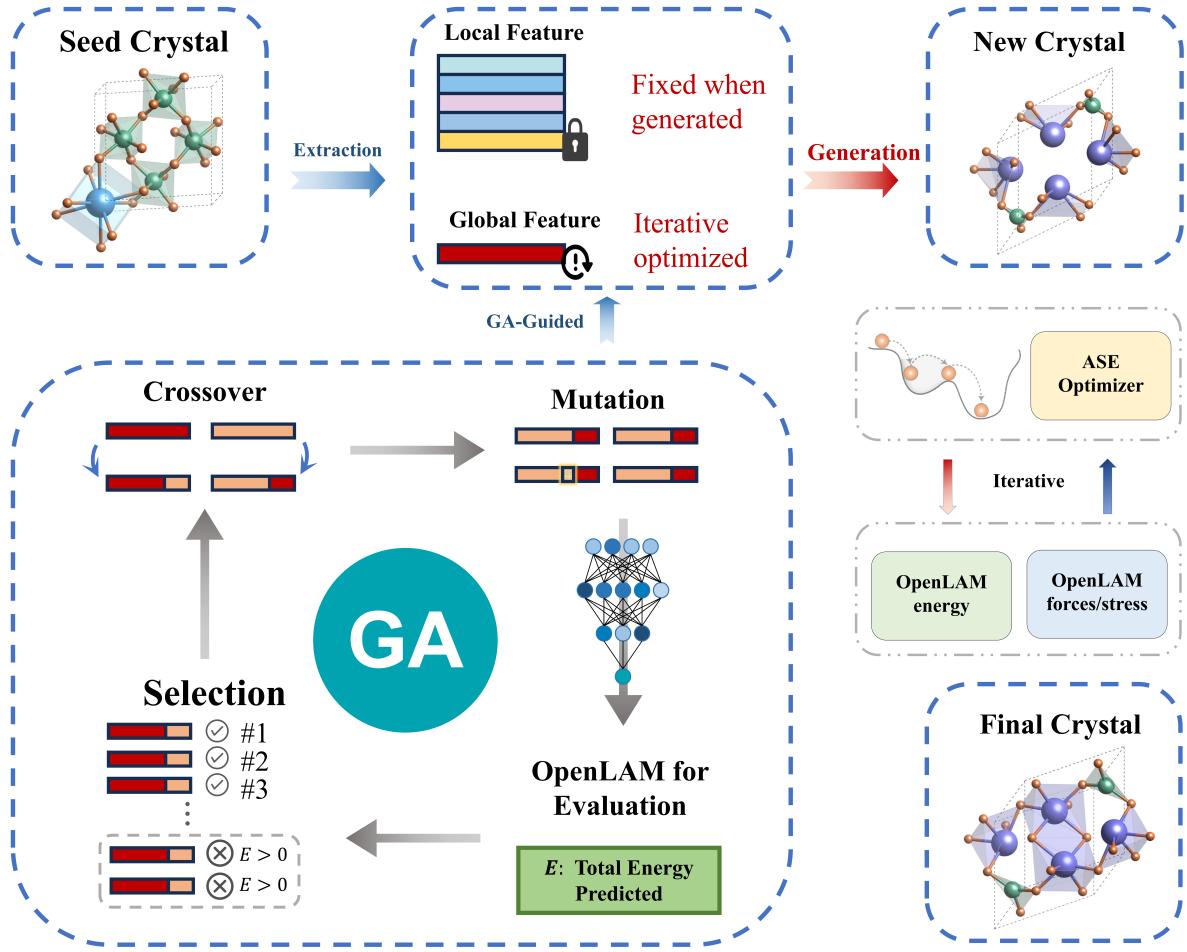


FIG. 2: The sampling process of VQCystal, involving fixed local latents, global latents sampling, and genetic algorithm optimization.

and CDVAE<sup>35</sup> require 22 seconds and 10 seconds, respectively. Thus, despite utilizing Transformers, VQCystal achieves highly competitive sampling times, balancing accuracy and speed effectively. Full analysis of sampling time is shown in the supplementary information.

### III. RESULTS AND DISCUSSIONS

#### A. Model performance on quantitative metrics

We evaluate the efficacy of our model on a diverse range of tasks to demonstrate its capability in generating high-quality structures of different crystals. Specifically, we focus on the training reconstruction indicators to evaluate representation learning task and focus on the validity and diversity of crystal generation task.

**Dataset:** We conduct experiments on three datasets: MP-20, Perov-5 and Carbon-24 following previous works<sup>30,35</sup>. The MP-20 dataset selects 45,231 stable inorganic materials from Material Projects<sup>21</sup>, including experimentally-generated materials with at most 20 atoms in a unit cell. The Perov-5 dataset<sup>22</sup> contains 18,928 perovskite materials with similar structures, each having 5 atoms in a unit cell. The Carbon-24 dataset<sup>23,24</sup> includes 10,153 carbon materials, with unit cells containing between 6 and 24 atoms. For all datasets, we follow a 60-20-20 train-validation-test split following previous works<sup>30,35</sup>. The details of the datasets can be found in supplementary information.

**Baselines:** We compare our model with two types of baselines. The first type includes deep learning based crystal generation models: FTCP<sup>16</sup>, Cond-DFC-VAE<sup>36</sup>, CDVAE<sup>35</sup> and our proposed VQCystal. These models generate new structures based on learned distributions. We also consider the second type including deep learning based crystal structure prediction (CSP) models: P-cG-SchNet<sup>37</sup> and DiffCSP<sup>30</sup>. Because CSP models can be adapted to do crystal generation task by randomly sampling the given crystal composition. This comprehensive comparison highlights the effectiveness of our VQCystal model in generating high-quality crystal structures.

### 1. *Representation Learning*

**Evaluation Metrics:** Following common practice, we evaluate by matching the predicted candidates with the ground-truth structure. The match rate is the proportion of matched structures over the test set. The matching process uses the StructureMatcher class in pymatgen<sup>38</sup> with thresholds stol = 0.5, angle\_tol = 10, and ltol = 0.3. The RMS is calculated between the ground truth and the best matching candidate, normalized by  $\sqrt[3]{V/N}$ , where  $V$  is the volume of the lattice, and averaged over the matched structures.

The results in Table I underscore the superior performance of our proposed VQCystal model compared to other deep generative models. For the Perov-5 dataset, VQCystal achieves a match rate of 95.60%, slightly lower than the best-performing baseline model, FTCP<sup>39</sup>, at 99.34%, but still a very high value, indicating an almost complete match. The RMS of VQCystal is 0.0438, which is well below 0.1Å, indicating that the differences are minimal and can be considered nearly identical in terms of actual crystal structures, despite the slight increase in RMS when compared to FTCP (0.0259). For the Carbon-24 dataset, VQCystal attains a match rate of 70.03%, surpassing both FTCP’s 62.28% and CDVAE’s 55.22%, with an RMS of 0.2573, indicating comparable performance in

TABLE I: Results on stable structure reconstruction task.

Model	MP-20		Perov-5		Carbon-24	
	Match rate↑	RMS↓	Match rate↑	RMS↓	Match rate↑	RMS↓
FTCP <sup>39</sup>	69.89	0.1593	99.34	0.0259	62.28	0.2563
Cond-DFC-VAE <sup>36</sup>	–	–	51.65	0.0217	–	–
CDVAE <sup>35</sup>	45.43	0.0356	97.52	0.156	55.22	0.1251
P-cG-SchNet <sup>37</sup>	15.39	0.3762	48.22	0.4179	17.29	0.3846
DiffCSP <sup>30</sup>	51.49	0.0631	52.02	0.0760	17.40	0.2759
VQCrystal	77.70	0.088	95.60	0.0438	70.03	0.2573

terms of structure accuracy. For the MP-20 dataset, VQCrystal achieves a match rate of 77.70%, outperforming both FTCP’s 69.89% and CDVAE’s 45.43%, while the RMS of 0.088, though not as low as CDVAE’s or DiffCSP’s, is still below 0.1Å—a very low value where the differences can be considered as minor internal variations within the crystal structure. These improvements highlight VQCrystal’s ability to capture the periodicity and discrete characteristics of crystal structures more effectively than other models. This can be explained by its use of discrete VQ to encode crystal structures, which aligns well with the inherent discrete nature of crystal lattices, providing a more accurate and effective representation.

## 2. Crystal Generation Task

To further evaluate the performance of our VQCrystal model, we conducted experiments on the crystal generation task. This task aims to generate new crystal structures that are valid, diverse, and have high coverage of the target space. We compare our model against several baseline methods using various metrics.

**Evaluation Metrics:** The results are evaluated from validity and diversity:

- **Validity:** We consider structural validity, compositional validity, energy validity, and force validity. The structural valid rate is calculated as the percentage of generated structures with all pairwise distances larger than 0.5Å, and the generated composition is valid if the entire charge is neutral as determined by SMACT<sup>40</sup>. Force validity are evaluated using the OpenLAM model with DeePMD-kit v2<sup>41</sup> as the DFT estimator. A structure is force valid if the maximum force  $f_{\max}$  is less than 0.05 eV/Å.

- **Diversity:** This metric evaluates how well the generated structures explore the space of possible crystal structures beyond those found in the original dataset. We use two metrics for this: Average Minimum Distance (AMD) and Fréchet Distance (FD)<sup>25</sup>. The AMD measures the average minimum distance between any generated structure and the ground truth structures using CrystalNN structural fingerprints<sup>42</sup> as input:

$$\text{AMD} = \frac{1}{|S_g|} \sum_{M_i \in S_g} \min_{M_j \in S_t} d_S(M_i, M_j)$$

The FD evaluates the distance between the distributions of generated and ground truth structures:

$$\text{FD} = \|\mu_g - \mu_t\|^2 + \text{Tr}(\Sigma_g + \Sigma_t - 2\sqrt{\Sigma_g \Sigma_t})$$

where  $\mu_g$  and  $\mu_t$  are the means, and  $\Sigma_g$  and  $\Sigma_t$  are the covariances of the generated and ground truth CrystalNN structural fingerprints<sup>42</sup> respectively.

TABLE II: Results on materials generation task.

Data	Method	Validity (%)			Diversity	
		Struc. $\uparrow$	Comp. $\uparrow$	Force $\uparrow$	AMD $\uparrow$	FD $\uparrow$
MP-20	FTCP <sup>37</sup>	1.55	48.37	-	-	-
	P-G-SchNet <sup>37</sup>	77.51	76.40	-	-	-
	CDVAE <sup>35</sup>	<u>99.98</u>	52.39	0.95	<b>0.165</b>	<u>0.132</u>
	DiffCSP <sup>30</sup>	99.94	<u>83.22</u>	<u>16.11</u>	0.099	0.025
	VQCrystal	<b>100.0</b>	<b>84.58</b>	<b>91.93</b>	<u>0.160</u>	<b>0.152</b>
Perov-5	FTCP <sup>37</sup>	0.24	54.24	-	-	-
	Cond-DFC-VAE <sup>36</sup>	73.60	82.95	-	-	-
	P-G-SchNet <sup>37</sup>	<u>79.63</u>	<b>99.13</b>	-	-	-
	CDVAE <sup>35</sup>	<b>100.0</b>	69.79	0.02	0.038	<u>0.025</u>
	DiffCSP <sup>30</sup>	<b>100.0</b>	<u>98.69</u>	<u>12.26</u>	<u>0.051</u>	0.002
	VQCrystal	<b>100.0</b>	97.48	<b>99.17</b>	<b>0.247</b>	<b>0.312</b>
Carbon-24	FTCP <sup>37</sup>	0.08	-	-	-	-
	P-G-SchNet <sup>37</sup>	48.39	-	-	-	-
	CDVAE <sup>35</sup>	<b>100.0</b>	-	0.00	<u>0.125</u>	<u>0.103</u>
	DiffCSP <sup>30</sup>	<b>100.0</b>	-	<u>0.01</u>	0.0187	0.033
	VQCrystal	<u>99.97</u>	-	<b>74.22</b>	<b>0.248</b>	<b>0.515</b>

The results in Table II show that VQCystal consistently achieves high validity rates across all datasets, with structural, compositional, and force validity metrics being significantly better than most baselines. Specifically, in the Perov-5 dataset, VQCystal reaches 100.0% in structural validity, 97.48% in compositional validity, and an impressive 99.17% in force validity, outperforming other models by a considerable margin. Because non-diffusion-based models like FTCP have very low structural and compositional validity, they do not calculate force validity. Although diffusion models are theoretically proven as mathematical frameworks for deep potential simulation, they still do not perform as well as our explicit optimization approach, with DiffCSP only achieving 12.26% force validity.

For the other datasets, similar trends are observed. On the MP-20 and Carbon-24 datasets, VQCystal demonstrates high force validity and other validity metrics, achieving a force validity of 91.93% on the MP-20 dataset and 74.22% on the Carbon-24 dataset, both significantly higher compared to other models.

The diversity metrics, e.g. AMD and FD, further validate the effectiveness of VQCystal, indicating that it can generate a more diverse set of crystal structures. Since diffusion models are known for their diversity, it is pertinent to compare VQCystal with diffusion-based methods like CDVAE. For the AMD metric, which measures average maximum deviation and indicates the spread of generated structures, VQCystal achieves 0.247 on Perov-5, 0.248 on Carbon-24, and 0.160 on MP-20. These values are higher than those achieved by CDVAE, which has an AMD of 0.038 on Perov-5, 0.125 on Carbon-24, and 0.165 on MP-20. Similarly, for the FD metric, which measures feature distance and indicates the distinctiveness of generated structures, VQCystal attains 0.312 on Perov-5, 0.515 on Carbon-24, and 0.152 on MP-20, compared to CDVAE’s 0.025 on Perov-5, 0.103 on Carbon-24, and 0.132 on MP-20. These results highlight VQCystal’s superior ability to produce a diverse and distinctive set of crystal structures. Overall, these findings underscore the capability of VQCystal in generating valid, diverse and well-covered crystal structures, making it a robust tool for crystal structure prediction and generation tasks.

## B. Interpretability

The sampling methodology of VQCystal is valid only if certain prerequisites are met. First, the global latent variable must contain rich information about the crystal to ensure meaningful variations. The local latent variables must retain enough information to keep the sampling process controllable and consistent with the original structure. Lastly, the

global latent space must be well-structured to help the genetic algorithm identify superior candidates. This section delves into the latent space and sampling process of VQCystal to validate these prerequisites.

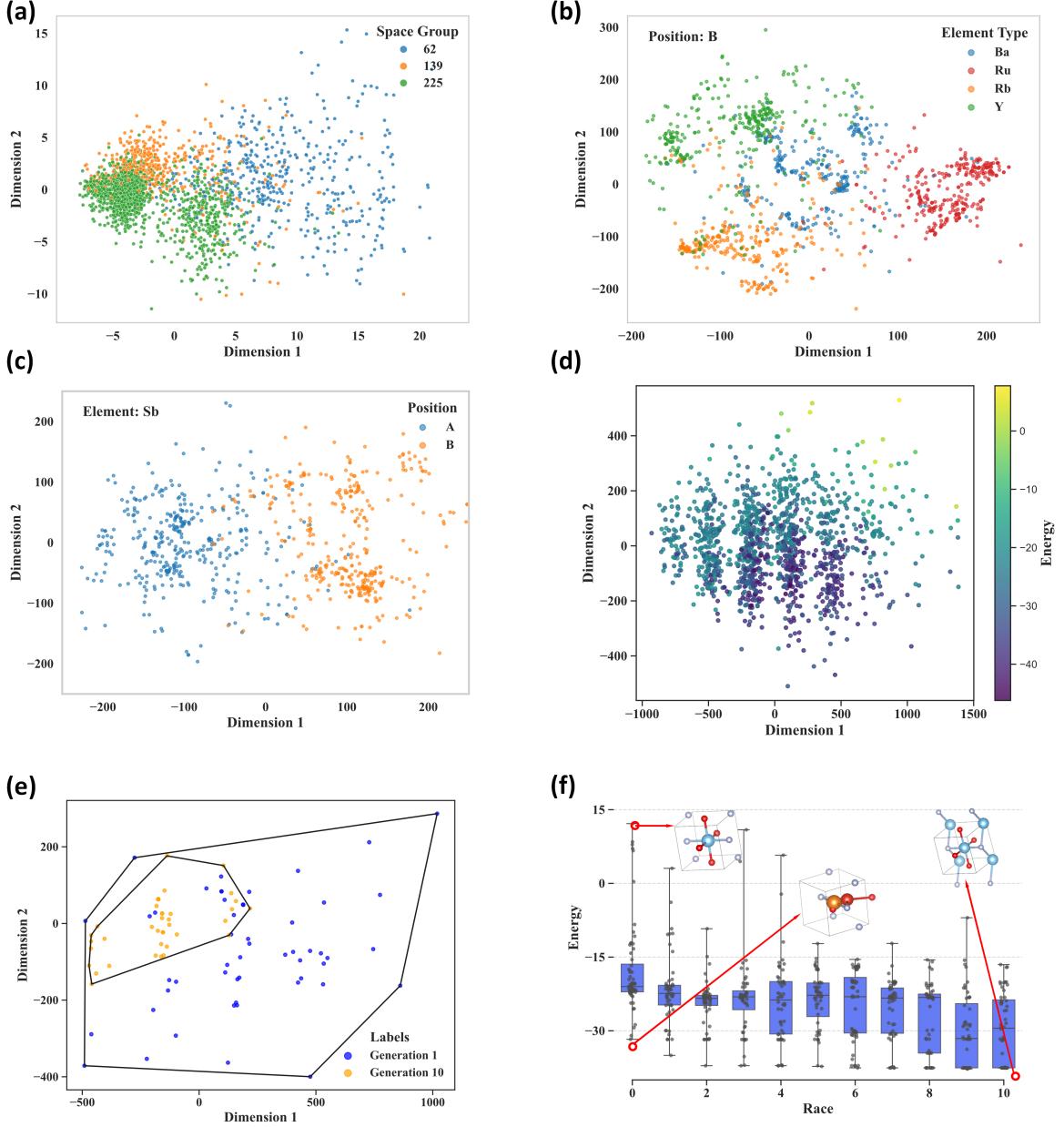


FIG. 3: Model interpretability analysis. (a, d) Analysis of global latent space. (b, c) Analysis of local latent space. (e, f) Analysis of sample space.

### *1. Global latent*

To substantiates the assertion that the global latents encapsulate comprehensive and abstract information, a crystal is randomly selected from the Perov-5 dataset, and its local latents are fixed. Subsequently, 1,000 different compositions of global indices are sampled and decoded into 1,000 crystal samples in conjunction with the fixed local latents. OpenLAM is utilized to estimate the total energy of these sampled crystals. The 1,000 decoded global latents, each of 128 dimensions, are projected onto two dimensions using Principal Component Analysis (PCA)<sup>43</sup>. The results are depicted in Figure 3(d), where each point corresponds to a global latent, color-coded based on the estimated energy of its respective sample. It is evident that the global latent space is well-organized by total energy, with high-energy regions smoothly transitioning to low-energy regions.

Further analyse aims to demonstrate that the global latent contains space group information. The MP-20 dataset is selected due to its rich diversity in space groups. 10,000 crystals are chosen, and their global latents are projected onto two dimensions using PCA. The data is then visualized based on their space group information in Figure 3(a). The three most frequently occurring space groups are selected for detailed analysis (space group:  $P6_2$ ,  $Fmmm$ ,  $Pm\bar{3}m$ ). A numerical analysis yields a Silhouette Score<sup>44</sup> of 0.478 with values above 0 generally suggesting that the data points are reasonably well-clustered. In this case, the high Silhouette Score indicates well-defined clusters, supporting that the global latent space captures space group information. This distinction is one aspect of the complex data in the global latent space, with more abstract details yet to be discovered.

### *2. Local latent*

To demonstrate the local latent variables contain sufficient atomic information, the Perov-5 dataset, consisting of perovskite materials with elements at ABX positions, was used. To check for element type information, the position need to be fixed and the B position together with four most common elements found at this position (Ba, Ru, Rb, Y) were selected. PCA was applied to reduce the dimensionality of the local latent to two components. The results were then plotted shown in Figure 3(b). Additionally, to demonstrate that local latent variables capture positional information, the most frequent elements found at both A and B positions were selected. Due to the constraints of the

ABX structure, where X is typically a non-metal and A and B are metals, only elements at A and B positions were visualized. Fixing the element types (Sb as it's one of the most frequent atom appearing in AB position) and clustering based on positions. The results in both Figure 3(b) and Figure 3(c) showed clear clustering of positions and types for these elements, reinforcing that the local latent variables effectively capture both atomic and positional information. The Silhouette Score was calculated to further quantify the clustering quality. The Silhouette Score of 0.2688 and 0.2518 indicates a moderate level of separation between clusters, with values above 0 generally suggesting that the data points are reasonably well-clustered.

### *3. The sampling process*

The previous section demonstrated that the global latent space is well-organized by the crystal's total energy (Figure 3(d)), thereby facilitating genetic algorithm searches. This subsection further validates the functionality of the genetic algorithm by examining each species during the evolutionary iterations. Figure 3(f) depicts the statistics of the total energy of the crystals across different iterations. These values were collected from a random genetic search process starting with a random initial crystal. Each point in Figure 3(f) represents a global latent in this evolutionary race, where the y-axis corresponds to the fitness function value during evolution, which is also the total energy of the decoded crystal as estimated by OpenLAM. A box plot for each race is provided to clearly illustrate the statistical distribution. It is evident that the overall total energy decreases as evolution progresses, with the 75th percentile dropping from approximately -21 to about -36. As the population iterates, the composition and lattice of the crystals gradually improve. Ultimately, the face-centered, body-centered, and specific point positions of the best sample in the final race are well-recognized by the VESTA software<sup>45</sup>.

Additionally, we projected the global latents onto two dimensions using PCA to visualize the evolutionary trends of the population. For clarity, Figure 3(e) only visualizes the initial and final generations. The black border represents the convex hull for the points of each generation, calculated using the scikit-learn package. It is clearly seen that the samples in a race gather and converge as evolution progresses. Combined with the aforementioned conclusions, this indicates that the genetic algorithm effectively captures the organized properties of the global latent space and gradually optimizes the species by simulating the evolutionary process.

## C. Reliable crystal generation of VQCrytsal

### 1. Design cases validated by Validation by datasets

To demonstrates the model’s capacity to generalize and reliably explore the search space of crystal structures. It is important to verify the model’s ability to generate materials that appear in the test set or the dataset but are not part of the training set.

For this purpose,VQCystal generates a large number of crystal structures, totaling 20,789 crystals trained on the MP-20 dataset. These crystals were first checked for duplicates within the training set. The ‘StructureMatcher’ class from the pymatgen library<sup>38</sup> was used to compare crystal structures and remove duplicates. After this process, 20,183 unique materials remained. These 20,183 materials were then compared against the complete Materials Project database as of June 1, 2018 with 132082 materials to check for duplicates with known structures in the dataset, excluding the training set. Among 2.16% were found to be duplicates of structures in the MP database, corresponding to 437 materials. Further analysis was performed by calculating RMS distances between the generated duplicates and their corresponding structures in the database. The average RMS distance for these duplicates was found to be 0.0509, with the smallest RMS distance reaching as low as 0.0001.

These results indicate that the duplicated structures are nearly identical to those in the database, validating the stability of the generated crystals. The close structural similarity to stable crystals in the MP database demonstrates that the generated materials are highly reliable and consistent with known stable structures.

### 2. Design cases validated by first-principles calculations

In addition to validation by datasets, to further demonstrate the reliability and practical applicability of the model, VQCystal is applied to specific design cases. Figure 4(a) illustrates the workflow for inversely designing of materials based on target properties. We selected materials with a bandgap ( $E_g$ ) between 0.5 and 2.5 eV, which is a desirable range for photovoltaic applications, and a formation energy ( $E_f$ ) less than -0.5 eV/atom to ensure chemical stability. We used models trained on the Materials Project database<sup>21</sup> to generate these materials. Firstly, The 20,789 crystals generated by VQCystal trained on the MP-20 data are screened to remove duplicates based on the complete Materials

Project database as of June 23, 2023<sup>21</sup>, leaving 19,776 unique structures. Following this, the structures are validated for compositional correctness using the SMACT library<sup>40</sup>, which checks for valid elemental combinations and charge neutrality. The validation process includes ensuring the elements' oxidation states and electronegativity are appropriate for forming stable compounds. This screening results in a 67.21% pass rate, reducing the number of valid crystal structures to 13,292. Following this, the structures' forces and energies are calculated using OpenLAM. Structures with an energy less than 0 and a maximum force ( $f_{\max}$ ) less than 0.05 eV/Å are considered stable. This step has a pass rate of 92.4%, resulting in 12,282 stable structures. Subsequently, structures with an excessive number of atoms, too many different types of elements, or containing lanthanide series elements are removed, leaving 2,771 materials.

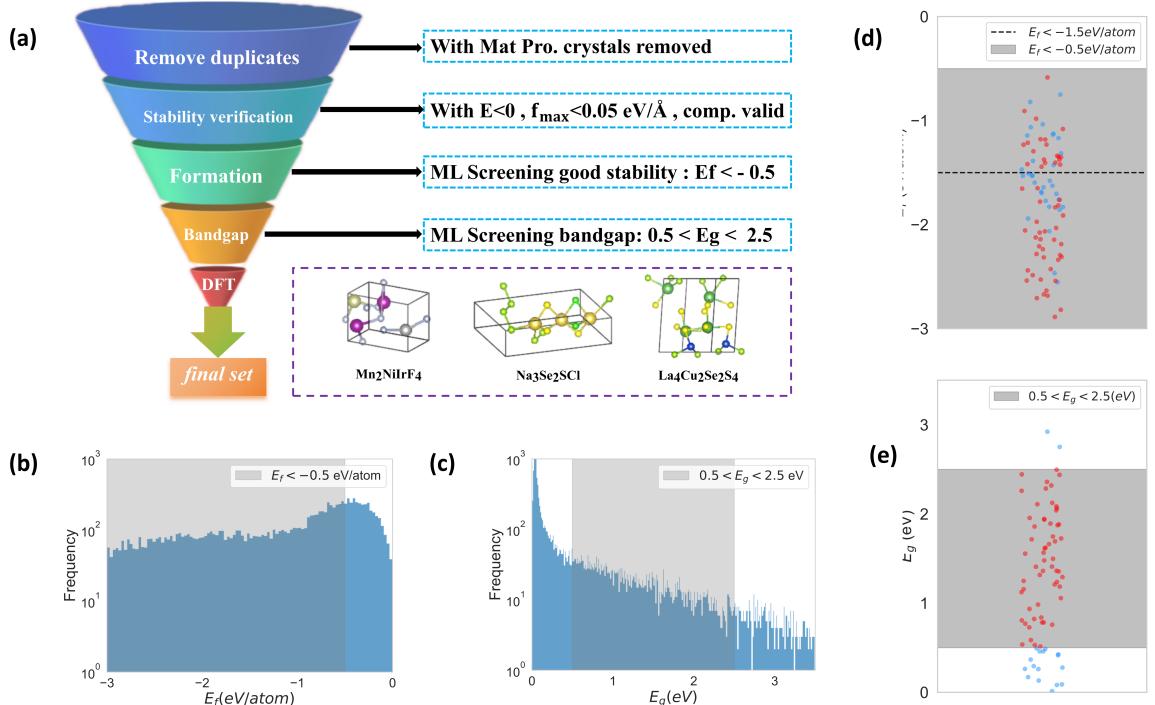


FIG. 4: (a) Workflow for designing materials based on target properties. (b) and (c) show the distribution of predicted formation energy and bandgap from the MEGNet model. (d) and (e) are scatter plots of the formation energy and bandgap of the 90 designed materials calculated using first-principles calculations.

Then, we employed the MEGNet model<sup>26</sup> and the ALIGNN model<sup>27</sup>, both of which are graph neural networks, to screen the target properties. The MEGNet model was trained on the Materials Project database as of June 1, 2018, achieving a mean absolute

error (MAE) of 0.028 eV/atom for formation energy and a test MAE of 0.33 eV for PBE bandgap. The ALIGNN model was trained on the JARVIS DFT dataset, with an MAE of 0.14 eV for the OPT bandgap and 0.033 eV/atom for formation energy. For the 2,771 materials, we predicted the bandgap and formation energy using both the MEGNet and ALIGNN regression models. Additionally, we used the MEGNet bandgap classifier to predict whether the bandgap is greater than 0 eV. We filtered the materials where the classifier predicted a bandgap greater than 0 eV and where all predicted properties fell within the specified ranges (bandgap between 0.5 and 2.5 eV, formation energy less than -0.5 eV/atom). This screening resulted in 92 materials. Figures 4(b) and 4(c) show the distributions of predicted formation energy and bandgap for the 12,282 materials before the removal of excessive elements, predicted using the MEGNet model. 90 out of 92 materials successfully passed the DFT relaxation process using the Vienna Ab initio Simulation Package (VASP)<sup>46</sup>. Next, we calculated the bandgap and formation energy of these 90 materials using first-principles calculations to validate the predictions.

The calculated bandgaps and formation energies were then compared with the target of bandgap ( $E_g$ ) between 0.5 and 2.5 eV, and formation energy ( $E_f$ ) less than -0.5 eV/atom. The results showed strong agreement in terms of formation energy, with 89 out of 90 materials having a formation energy lower than -0.5 eV/atom, indicating an almost 100% hit rate for stability prediction, which strongly suggests the stability of the predicted crystals. Regarding the bandgap, 56 out of the 90 materials had a bandgap within the target range of 0.5 to 2.5 eV. Figure 4(d,e) shows the distribution of the designed 90 materials, while the red points indicates materials which hit the target of both  $E_f$  and  $E_g$ . The band structures are shown in supplementary materials. These results further confirmed the reliability of the VQCystal model in predicting material properties, particularly for formation energy, and demonstrated the effectiveness of using machine learning models like MEGNet and ALIGNN for large-scale material discovery. Details of the materials is shown in the supplementary information.

### 3. Generation cases of 2D materials

Two-dimensional (2D) materials have gained significant attention due to their unique physical and chemical properties, which offer promising applications in areas such as energy, electronics, and catalysis. Compared to traditional three-dimensional materials, 2D materials feature an ultra-thin structure and high surface area, allowing for excep-

tional electrical, optical, and mechanical behaviors<sup>47</sup>. Building upon the significance of 2D materials, we applied VQCystal to generate new structures from a comprehensive 2D materials database called C2DB<sup>28,48</sup>. Specifically, we trained a VQCystal model using the C2DB database, which contains a total of 3,521 2D materials. The VQCystal model was trained with a 60-20-20 train-validation-test split. On the validation set, the model achieved a match rate of 88.56%.

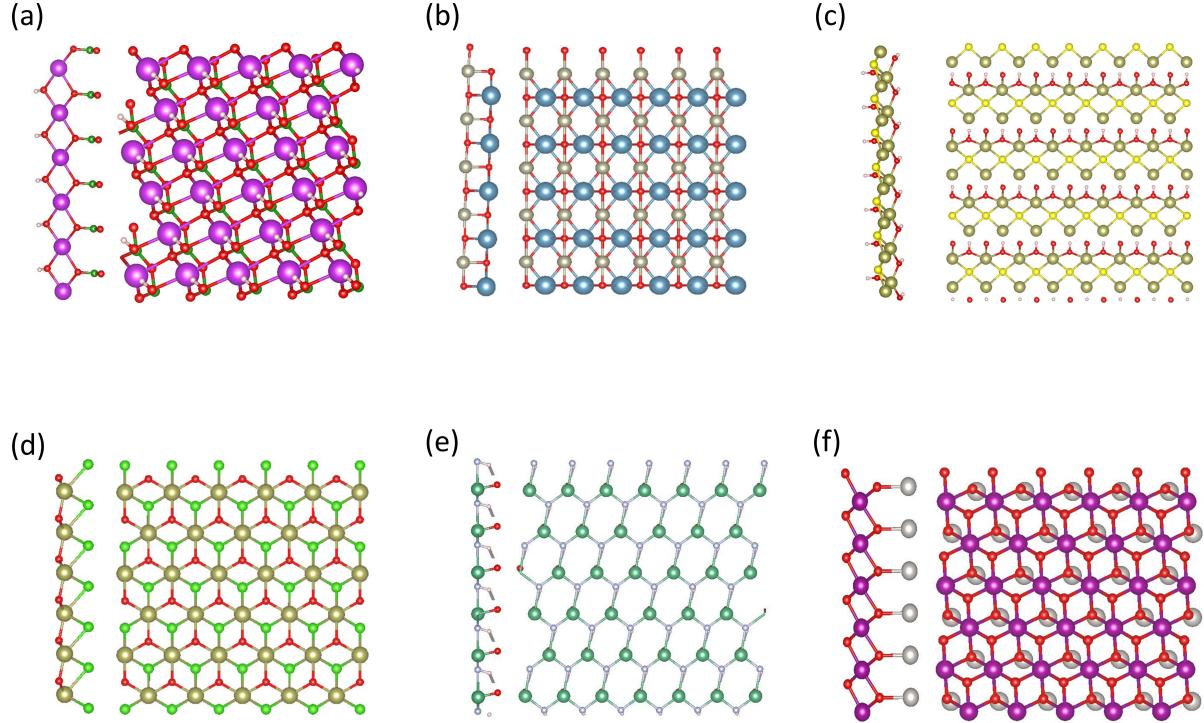


FIG. 5: (a-f) Generated two-dimensional materials along with their formation energy and total energy.

Afterward, we generated nearly 12,000 candidate materials, which were then filtered using the Atomic Simulation Recipes<sup>49</sup> to determine whether they were truly two-dimensional. This process left us with 3,521 materials. Subsequently, we removed duplicates by comparing the generated materials with the C2DB database, utilizing the ‘StructureMatcher’ class from the pymatgen library with the parameters  $\text{ltol}=0.3$ ,  $\text{stol}=0.5$ , and  $\text{angletol}=10$ . After this deduplication step, 2,638 candidate materials remained. Following the deduplication, we applied similar filtering steps as previously described, ensuring that the materials met criteria for elemental composition, structural validity, and reasonable force values. In addition, lanthanides, actinides, and structures containing an excessive number of different elements were removed. After this filtering process, 846 candidate materials remained. To assess the stability of these materials,

MEGNet and ALIGNN models were used to predict the formation energy of the 846 materials, following the same procedure as in the inverse design of the MP-20 datasets. Materials with a predicted formation energy less than -0.5 eV/atom in both models were selected, resulting in 184 materials. A random selection of 26 materials from the filtered set underwent DFT relaxation, with 23 successfully passing the process. Further calculation of their formation energy. Out of the 23 materials, 19 exhibited a formation energy lower than -0.5 eV/atom, with 17 of them, representing 73.91%, having a formation energy lower than -1 eV/atom. This indicates a good level of stability for the generated two-dimensional materials. Figure 5 presents a selection of the generated two-dimensional materials, illustrating both their formation energy and total energy. These visual representations emphasize the effectiveness of the VQCystal model in generating stable 2D materials, further supported by the energy analysis. Further information about the generated materials is shown in the supplementary materials.

#### IV. CONCLUSION

In summary, in this study we introduce VQCystal, an innovative pipeline for crystalline materials discovery that integrates a hierarchical VQ-VAE representation learning module, the open-source machine learning-based structural relaxation method Open-LAM, and a genetic algorithm. The effectiveness of incorporating discreteness in representation learning is demonstrated through benchmark performance on diverse datasets. Furthermore, metric analysis of the sampled novel crystals reveals that the VQCystal pipeline successfully discovers both valid and diverse novel crystalline materials. Interpretation of the results indicates that VQCystal has developed a highly interpretable latent space at both global and atomic levels. For inverse design tasks, we employed a genetic algorithm to search for stable candidates, followed by a series of filtering processes. In the case of 3D crystals, DFT validation confirmed that 50% of bandgaps and 99% of formation energies of the 56 filtered materials matched the target ranges of bandgap between 0.5 and 2.5 eV and formation energy below -0.5 eV/atom. For 2D crystals, DFT validation revealed that 73.91% of 23 filtered structures exhibited high stability, with formation energies below -1 eV/atom.

## DATA AND CODE AVAILABILITY

The Perov-5, Carbon-24, MP-20 datasets are queried from cdvae<sup>17</sup> at <https://github.com/tmse93/cdvae>. The Materials Project dataset is queried from its website<sup>21</sup> in June, 2023. (Note a query with the same criteria now would yield a different number of crystals from the recorded number in the study due to the updates and the addition of crystals of the Materials Project.) The C2DB<sup>48</sup> database is required from the database of jarvis-tools<sup>50</sup>.

---

\* These two authors contributed equally to this work.

† cenyang@fudan.edu.cn

‡ zhangh@fudan.edu.cn

<sup>1</sup> Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

<sup>2</sup> Seiji Kajita, Nobuko Ohba, Ryosuke Jinnouchi, and Ryoji Asahi. A universal 3d voxel descriptor for solid-state material informatics with deep convolutional neural networks. *Scientific reports*, 7(1):16991, 2017.

<sup>3</sup> Vadim Korolev, Artem Mitrofanov, Artem Eliseev, and Valery Tkachenko. Machine-learning-assisted search for functional materials over extended chemical space. *Materials Horizons*, 7(10):2710–2718, 2020.

<sup>4</sup> Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

<sup>5</sup> Edward O Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annual Review of Materials Research*, 45(1):195–216, 2015.

<sup>6</sup> Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

<sup>7</sup> Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

- <sup>8</sup> Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.
- <sup>9</sup> Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Gutfleisch, and Hongbin Zhang. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):66, 2021.
- <sup>10</sup> Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.
- <sup>11</sup> Yong Zhao, Edirisuriya M Dilanga Siriwardane, Zhenyao Wu, Nihang Fu, Mohammed Al-Fahdi, Ming Hu, and Jianjun Hu. Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Computational Materials*, 9(1):38, 2023.
- <sup>12</sup> Zhipeng Li and Nick Birbilis. Nsgan: a non-dominant sorting optimisation-based generative adversarial design framework for alloy discovery. *npj Computational Materials*, 10(1):112, 2024.
- <sup>13</sup> Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- <sup>14</sup> Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- <sup>15</sup> Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- <sup>16</sup> Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.
- <sup>17</sup> Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

- <sup>18</sup> Hans Wondratschek and Ulrich Muller. International tables for crystallography, symmetry relations between space groups. *International Tables for Crystallography*, 5(5):732–740, 2004.
- <sup>19</sup> T. Hahn. International tables for crystallography, volume a: Space group symmetry. *Published for the International Union of Crystallo*, 1987.
- <sup>20</sup> deepmodeling. Openlam. <https://github.com/deepmodeling/openlam>, 2024. Accessed: 2024-08-03.
- <sup>21</sup> A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. The materials project: a materials genome approach to accelerating materials innovation. *apl mater* 1: 011002, 2013.
- <sup>22</sup> Ivano E Castelli, David D Landis, Kristian S Thygesen, Søren Dahl, Ib Chorkendorff, Thomas F Jaramillo, and Karsten W Jacobsen. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy & Environmental Science*, 5(10):9034–9043, 2012.
- <sup>23</sup> Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.
- <sup>24</sup> Chris J Pickard and RJ Needs. High-pressure phases of silane. *Physical review letters*, 97(4):045504, 2006.
- <sup>25</sup> Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. 1994.
- <sup>26</sup> Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.
- <sup>27</sup> Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.
- <sup>28</sup> Sten Haastrup, Mikkel Strange, Mohnish Pandey, Thorsten Deilmann, Per S Schmidt, Nicki F Hinsche, Morten N Gjerding, Daniele Torelli, Peter M Larsen, Anders C Riis-Jensen, et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials*, 5(4):042002, 2018.
- <sup>29</sup> Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- <sup>30</sup> Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information*

- <sup>31</sup> Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.
- <sup>32</sup> Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex volutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.
- <sup>33</sup> Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. Crystal structure prediction by combining graph network and optimization algorithm. *Nature communications*, 13(1):1492, 2022.
- <sup>34</sup> Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.
- <sup>35</sup> T Xie, X Fu, OE Ganea, R Barzilay, and T Jaakkola. Crystal diffusion variational autoencoder for periodic material generation, 2021. URL <https://arxiv.org/abs/2110.06197>, 2110, 2021.
- <sup>36</sup> Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, 2020.
- <sup>37</sup> Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.
- <sup>38</sup> Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics ( pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.
- <sup>39</sup> Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.
- <sup>40</sup> Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.
- <sup>41</sup> Jinzhe Zeng, Duo Zhang, Denghui Lu, Pinghui Mo, Zeyu Li, Yixiao Chen, Marián Ryník, Ling Huang, Ziying Li, Shaochen Shi, et al. Deepmd-kit v2: A software package for deep

- potential models. *The Journal of Chemical Physics*, 159(5), 2023.
- <sup>42</sup> Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10(10):6063–6081, 2020.
- <sup>43</sup> Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
- <sup>44</sup> Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.
- <sup>45</sup> Koichi Momma and Fujio Izumi. Vesta: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied crystallography*, 41(3):653–658, 2008.
- <sup>46</sup> G. Kresse A and J. Furthmller b. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set - sciencedirect. *Computational Materials Science*, 6(1):15–50, 1996.
- <sup>47</sup> Hua Zhang. Introduction: 2d materials chemistry, 2018.
- <sup>48</sup> Morten Niklas Gjerding, Alireza Taghizadeh, Asbjørn Rasmussen, Sajid Ali, Fabian Bertoldo, Thorsten Deilmann, Nikolaj Rørbæk Knøsgaard, Mads Kruse, Ask Hjorth Larsen, Simone Manti, et al. Recent progress of the computational 2d materials database (c2db). *2D Materials*, 8(4):044002, 2021.
- <sup>49</sup> Morten Gjerding, Thorbjørn Skovhus, Asbjørn Rasmussen, Fabian Bertoldo, Ask Hjorth Larsen, Jens Jørgen Mortensen, and Kristian Sommer Thygesen. Atomic simulation recipes: A python framework and library for automated workflows. *Computational Materials Science*, 199:110731, 2021.
- <sup>50</sup> Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.