# A scalable crystal representation for reverse engineering of novel inorganic materials using deep generative models

Rochan Bajpai[1#], Atharva Shukla[2], Janish Kumar[2], Abhishek Tewari[1,3*]

1) *Department of Metallurgical and Materials Engineering, Indian Institute of Technology Roorkee, Uttarakhand, India*
2) *Department of Physics, Indian Institute of Technology Roorkee, Uttarakhand, India*
3) *Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Uttarakhand, India.*

**Corresponding author email: abhishek@mt.iitr.ac.in*

*[#]Current address: Department of Materials Science and Engineering, Carnegie Mellon University, USA.*

**Abstract**

The efficient search for crystals with targeted properties is a significant challenge in materials discovery. The rapidly growing field of materials informatics has so far primarily focused on the application of AI/ML models to predict the properties of known crystals from their fundamental and derived properties as descriptors. In the last few years, deep learning-based approaches have spawned a slew of innovative data-driven materials research applications. Materials scientists have used these techniques for the reverse engineering of crystal structures for target applications. However, one of the challenges has been the representation of the crystal structures in the machine

readable format. Proposed representations in the literature lack in generality and scalability. In this paper, we train a conditional variational autoencoder with a scalable and invertible representation along with the elemental properties of the constituents as descriptors to inverse-design new crystal structures with specified attributes. When targeting formation energy, we show that our model predicts structures that are not in the complete OQMD database. Finally, we use first-principles density functional theory calculations to validate our findings and show that the developed model is able to generate novel crystal structures for targeted property, i.e. formation energy in this case.

## 1. Introduction

The community of material scientists has long been interested in the hunt for novel crystal structures with promising properties. Indeed, solutions to today's grand challenges hinge on innovations in materials: photovoltaics, catalysis, drug delivery, energy storage and more. However, the need for new materials exceeds the conventional materials discovery process's ability to produce them. When discovering novel materials with remarkable properties, we rely heavily on serendipity.

Though viewed as the ultimate aim of the materials community, inverse design is still a long way off. Solid-state materials design has always been driven by experimental study and scientific intuition. The traditional materials discovery process involves "trial-and-error" experimentation, which, given the vastness of the materials space, is not feasible if one wants to look for novel structures with specific properties rapidly.

First-principles and high-throughput computational techniques for materials discovery are now feasible because of significant advances in computing power. Large data sets of materials, e.g. AFLOW library[1], Open Quantum Materials Database (OQMD)[2], and Materials Project database[3],

have spawned from running these calculations on millions of possible structures and stoichiometries. These databases generally list crystal characteristics and crystal shapes as established by density functional theory (DFT). However, the enormous computational overhead required to run these calculations on every possible crystal structure suggests a need for more efficient methods to traverse the materials space.

A natural alternative to this approach is making use of existing machine learning techniques, which have already seen success in several applications. The goal of ML approaches employed in the context of materials is to learn a function that maps a material to the desired attribute. DL techniques that can predict the underlying probability distribution of both structure and property and connect them are known as deep generative models. These models may extract prominent properties that define crystals by making use of trends in large datasets.

Recently, existing generative models have been adapted to discover previously unknown stoichiometries, molecules[4], and solid crystals[5–7], with the latter still in their nascent stages of development. Work done in generating new crystals have either had restricted stoichiometry or structures. In contrast to molecules, generative modelling of inorganic solid-state materials has been challenging because of the lack of invertible representation and limited availability of datasets. Attempts have been made to overcome these obstacles. Some models have alleviated these problems and presented a general framework for the generation of crystals. Nevertheless, we believe they lack in creating a model, which can be generalized to the vast variety of compositions possible out of the periodic table and scalable to generate the inorganic materials with complex chemistries.

In the space of generative modelling of inorganic crystal structures, studies can be differentiated primarily based on two aspects: crystal structure representation and learning models used for

generating the novel crystal structures. One of the earliest works in this domain was the rediscovery of experimentally known $V_xO_y$ materials when the model was trained without them. In addition to this, 20,000 hypothetical compounds were generated as well. The **image based representation**[8] used in this work was named iMatGen, where the crystal structure was encoded using two images, cell (length of the cell edges and angles between them) and basis (atomic position within a unit cell). This image based representation was encoded and then decoded to original data using a Variational Autoencoder.

In addition to the use of variational autoencoders, generative adversarial networks (GAN) have also been employed for the generation of material crystals. The focus in the work done by Kim et al.[9] was on the generation of Mg-Mn-O ternary crystals using a **point cloud representation** which was a combination of unit cell parameters and fractional coordinates of the elements present in the crystals. These are concatenated to create a 2D matrix. This is useful as it has less memory requirement (by a factor of 400) in comparison to the iMatGen. Data augmentation was used to overcome the lack of translational, rotational and supercell invariances. The representation used in this work was similar to the one we introduce in the current work. But it differs in the sense that it was constrained by the composition of the materials which could be generated, e.g. model developed by Kim et al. could be used only for generating different stoichiometries of Mg-Mn-O ternary crystals.

Pathak et al.[10] used the **one-hot key representation** for generating inorganic crystal structures using conditional variational auto-encoders (CVAE) and deep neural network to predict three properties of the generated materials namely, enthalpy of formation, volume per atom and energy per atom. However the representation contained only compositional information without providing any 3-d structural information, i.e. the target crystal structure needs to be fixed initially. In a recent

study, Turk and coworkers[11] used a simplified version of this representation to generate stable Elpasolite compositions and compared three generative models; reinforcement learning, VAE and GAN. They concluded that although all the three models are capable of creating novel crystal structure, VAE and GAN are more reliable in terms of reproducibility.

Another kind of representation used for a General Adversarial Network is the translation of lattice constants and atomic positions into a **voxel space**, followed by encoding into a 2D crystal graph using an autoencoder. This helps change the heterogenous and discontinuous representation in the CIF files to a continuous and homogeneous representation. Long et al.[12] used this representation to create stable crystal structures of multicomponent systems based on their formation energy. Court et al.[13] also used voxelized electron density maps to represent the crystal structure. Formation energy value was also concatenated along with the electron density maps. They used VAE along with GAN to generate crystal structures of different classes of materials. However, they needed to use a combination of UNet semantic segmentation network and morphological transformations to convert this representation back to a crystal structure.

We note that the work done by Ren et al.[14] is similar to what we discuss in this paper. Their representation used the information in the CIF file to create a **2D input matrix**. In addition, they also include the elemental properties by projecting them to different crystal planes using a discrete fourier transform. Similar to us they also use an autoencoder based model for learning the underlying distribution of the dataset and generating new crystal structures. The difference lies in our generation of crystals with some targeted properties. Their work makes use of local perturbation (a sampling technique) to generate crystals similar to the ones already present in the dataset. Whereas our model is able to randomly generate crystals with the desired property by sampling out from the appropriate dimension in the latent space. Furthermore, we note that our

approach allows for greater flexibility for incorporating new sites and elements as the Ren et al.[14] representation is currently limited up to ternary compounds.

To summarize, while crystal graphs provide higher flexibility in handling input crystal size compared to other representations, their inherent non-invertibility[15] renders them unsuitable for inverse design applications. Alternative representations like voxelized electron density maps are currently constrained to handling cubic structures, whereas iMatGen is memory intensive relative to 2D representations. In this work, we develop a new representation compatible with a CVAE, which puts no restrictions on the type of elements or the crystal structure. We use a CVAE and train it on our dataset to create a multidimensional latent space, which also has the values of formation energy encoded in one of its dimensions. The use of CVAE helps us generate new targeted stable or unstable crystals. We verify the results of CVAE using a neural network, which predicts the formation energy of the generated crystal structures. This is done to reduce the dependence on computationally expensive DFT calculations. In the end, we also use DFT calculations on a subset of our generated crystals to verify our results.

## 2. Computational Method

### 2.1. Dataset

We used the open-source OQMD database to train our CVAE model. Compounds having a maximum of four distinct elements and fewer than six atoms in the unit cell were considered. They were also required to have an OQMD Stability of less than 0.1, which allowed us to restrict our search to potentially synthesizable materials. Finally, compounds having negative formation energy were regarded as stable in conditional generation for formation energy. There are 48707

data points in the database, comprising 44474 stable and 4233 unstable structures. The data was stratified-split into 37600 and 11007 training and validation data points, respectively.

### 2.2. Representation

The representation employed in generative modelling must be invertible, i.e. the output of the neural network model must immediately convert to a unique unit cell. The representation should also include the crystal's crystallographic and stoichiometric information as well as the attributes of the constituent elements. Keeping in view these points, the representation we used went through numerous iterations before we narrowed it down to the final representation based on various tradeoffs that are discussed in this section.

To create the first representation, basic information was taken from the crystallographic information file (cif). The length and angle of the three translation vectors were scaled to a maximum value of 1 by dividing them by 10 and 180 degrees, respectively, giving us a (2,3) matrix. This matrix was padded with zeros to make it a (2,5) matrix. These make up the first two rows of our (7,5) dimensional matrix. The rest of the rows described each atomic site present in the unit cell. The first column stores the period of the element, followed by the group. The last three columns give us the fractional coordinates of the element. The rows were zero-padded if the number of sites is less than 5. We call this simple representation *Rep1* and use it as a benchmark. *Rep1* contained no information about the properties of the elements present in the material. Due to this, it was difficult for our model to learn the properties of the compound, which depend on the elemental properties, e.g. formation energy, as discussed in section 3.1. Therefore, the properties of the elements present at each site were also added to the representation. The elemental properties were one-hot encoded. When working with generative models, it is better to have dimensions

whose interdependencies are not complex, which is why we believe one hot encoding makes the process of learning easier for the encoder.

For describing the elemental properties at each site, we use two separate property sets. First, we use the one-hot encoded elemental properties from the crystal graph convolutional neural network (CGCNN)[16] to give us a row vector of 138 columns. The properties used are; electronegativity, covalent radius, valence electrons, first ionization energy, electron affinity, block and atomic volume. As shown in figure 1, the first row contains the unit cell information and the last row of the representation signifies whether the structure is stable (1) or not (0). In between the rows contain the information of the lattice sites, i.e. the group, period, fractional coordinates and elemental properties. Thus, a matrix of (N+2,138) size was created which contains the information of the crystal structure as well as the elemental properties. N is the maximum number of sites present in a unit cell in the entire dataset. This is referred to as *Rep2*.

A second set of 63 elemental properties from the Magpie[17] elemental descriptors were used to create *Rep3*. We concatenate these 63 properties of each element present in the crystal to our original representation describing the structure to get a 2D matrix representation of size (N+2,68). Since the number of elemental properties is much larger in Rep3 in comparison to Rep2, we do not bin the data to keep the dimensionality of the input computationally manageable. It is notable that all of the three representations have no constraint on a particular geometry, stoichiometry or the number of sites.

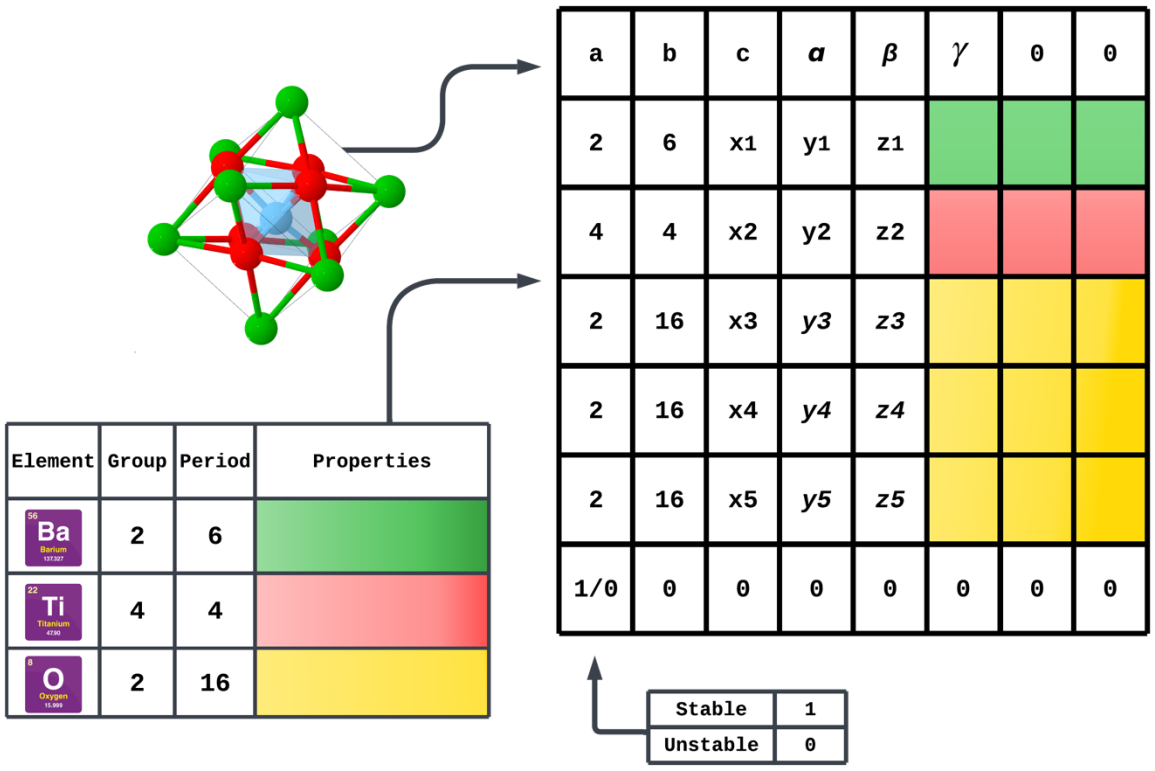


**Figure 1.** *A 2D matrix of BaTiO3 crystal structure with 5 sites. First row consists of cell parameters and the last row first column gives the information whether the structure is stable (1) or unstable (0). The size of the resulting representation for a general crystal with N sites will be (N+2, i+5) where N is the number of sites and i is the size of the elemental properties vector.*

The fractional coordinates of the atomic sites were also one-hot encoded. This is done based on the observation that the fractional coordinates present in the already known crystal structures can be separated into distinct bins. The combinations of the three fractional sites (x,y,z) can also be grouped together based on the frequency of their occurrence. Since the values in our input representation of the unit cell exist on a continuous spectrum determined by the scale and units of the relevant parameters, we first need to round off the values in order to bin them. However,

rounding them off to some arbitrary values would include unwanted errors by misplacing commonly found values in our representation. To deal with this, we bin the different parameters around the values which frequently appear in the distribution of that parameter in the whole dataset. We choose the number of bins equal to the number of peaks in the distribution for a parameter. And then, every parameter value is shifted to the value corresponding to the closest peak in the frequency distribution. As an example, rounding off to the nearest 0.1 fractional coordinate would convert 0.125, a very common coordinate to 0.1 (fig. 2a), which is not common in our system (fig. 2c). However, our method would keep 0.125 as it is and instead round off nearby sites to that number (fig. 2b). The site coordinates of the atoms are one hot-encoded after the rounding-off process.

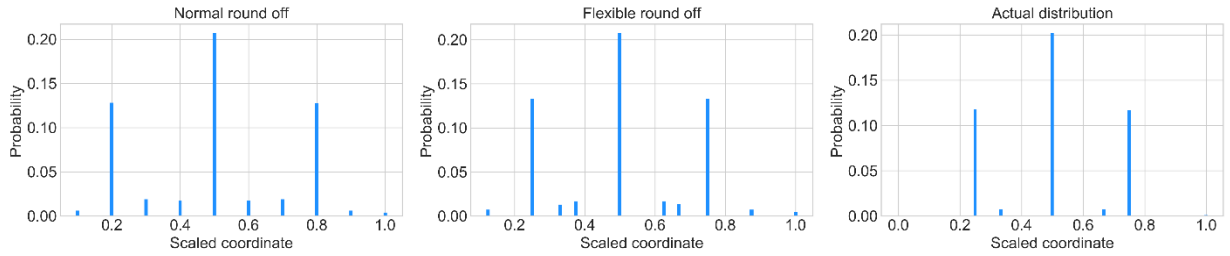


**Figure 2.** *Comparison between fractional coordinates distributions for different round off techniques. a) rounding off to the nearest 0.1 fractional, b) rounding off to 11 discrete values based on the frequency of occurrence, and c) rounding off to 4 decimal places.*

When comparing our final representation with a similar 2D representation proposed by Ren et al.[14], we emphasize that our approach allows for greater flexibility for incorporating new sites and elements, i.e. we just need to add a new row in the 2-D input matrix. The addition of more elements is more significant of the two as their representation is currently limited up to ternary compounds. Addition of new sites is not straightforward in their representation.

### 2.3. Formation Energy Model

To select the best representation to train our generative model, a formation energy model was built and trained on each of the representations and the results were compared. Additionally, the results were also compared with the CGCNN model to test the predictive accuracy of the representations. The model was also used later for validating the generated stable/unstable crystal structures.

The formation energy model is a fully connected neural network with residual layers. The model consists of 7 blocks of layers with each block consisting of a dense layer followed by a batch normalization layer, a LeakyRelu layer and a dropout layer. The seventh block does not have a dropout layer and is connected to a single neuron, which outputs the formation energy. The input to each block is the output of the previous block combined with the original input. Mean absolute error (MAE) was used as the loss metric to assess the performance of the model. Adam optimizer was used with default parameters.

### 2.4. Conditional Variational Autoencoder

To generate new crystal structures, we use a Variational Autoencoder (VAE), a generative model. VAEs have already shown promise in generating many kinds of complicated data, including handwritten digits, faces, physical models of scenes, segmentation, and predicting the future from static images. VAEs have also been used previously to generate crystal structures[10,11,14]. They also give us the freedom to use different sampling techniques, which can be of help when trying to generate crystals similar to the ones already known[14], as the latent space near already encoded crystals are bound to have crystals with similar properties. The goal of the current work was generation of novel materials with a targeted property (here formation energy) via a purely unsupervised design method. To achieve this, we extend the VAE to a Conditional Variational

Autoencoder (CVAE) and use our already existing database to label the crystals with formation energy greater than 0 and those smaller than 0.

In order to generate crystals with a specific target property, the latent space of a VAE must be structured accordingly. In our approach, we employ a CVAE that segregates the dimensions of our crystals based on the target property. This methodology allows us to generate crystals with a high probability of possessing the desired property without need for local perturbation, which has the drawback of generating structures similar in properties and structure to the crystal perturbed. This represents a promising initial step towards materials discovery with minimal user input, as it potentially enables us to obtain results through random sampling alone. The generation process flowchart utilizing CVAE is depicted in Figure 3. The various components of the CVAE model are discussed in the subsequent sections.

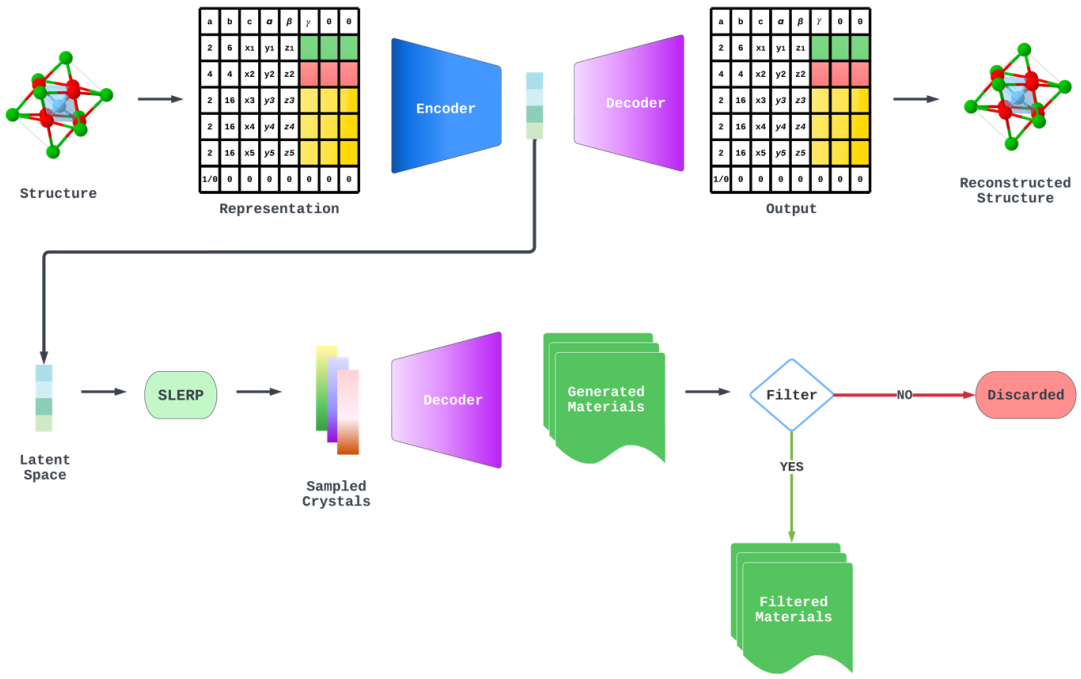


**Figure 3.** *Flowchart of the crystal structure generation process using conditional variational autoencoder.*

### 2.4.1. Encoder

The encoder encodes the information present in our representation in a lower-dimensional latent space. This latent space is structured into different regions for the stable and unstable crystals. We use a 1D convolutional neural network (CNN) to reduce the dimensionality of our input and map it to a latent space. This approach is followed to capture the spatial dependence of our crystal representation. Similar approaches have been used before for 3D image classification as well as in materials space[18]. Two layers of 1D CNN of kernel size (5,3), strides (1,2), and padding (2,1) were used. The number of channels are 16 and 32, respectively. After the convolutional layer, we flatten the array to output 1024 neurons. This is further reduced to size 100 to optimize the performance, which is the dimensionality of our latent space, before being upscaled back to size 2208. We use leakyRELU activation layer (parameter = 0.2). Batch Normalization is used between the layers.

### 2.4.2. Decoder

The encoded points in the latent space act as an input to our decoder which aims to recreate the corresponding crystal structure. The decoded output contains all the information present in the CIF file and hence, corresponds to a unique crystal structure. According to our hypothesis, a point sampled out from the "stable" region of our latent space should generate a stable compound. Two transpose conv1D layers with kernel size (3,4), padding (1,1), stride (2,1) and LeakyRELU (0.2) activation layer were used.

### 2.4.3. Loss

The loss function consists of the reconstruction loss, which is modelled by mean squared error and the KL-divergence loss. Both these losses are scaled with different parameters. The choice of these functions is standard for a VAE. We note that the dice loss can also be used for parts of our matrix

given the binary nature of our input. The difference in performance was negligible. MSE is scaled by a factor of 28.8 and KLD is scaled by a factor of 1.13. ADAM optimizer was used with an initial learning rate of 0.0005 and the inbuilt scheduler ReduceLROnPlateau with patience = 3 was also used. We train our model for 200 epochs.

**2.5. Crystal Structure Generation and Filtering**

We use spherical linear interpolation (SLERP[19]) method to sample points from our latent space in order to discover new materials. This is a sound way to interpolate between two n-dimensional vectors,

$$Slerp(q_1, q_2; \boldsymbol{\mu}) = \frac{\sin\ (1-\mu)\theta}{\sin\theta}\, q_1 + \frac{\sin\mu\theta}{\sin\theta}\, q_2$$

Where, μ is the weight assigned to $q_2$, (1-μ) is the weight for $q_1$ and θ is the angle between $q_1$ and $q_2$.

To test the efficacy of our conditional generation strategy, we sample points from both the constructed stable and unstable zones. Crystals that have the same stoichiometry as those in our test or train set are referred to as rediscovered. We run the produced structures via the SMACT filter described in reference[20] to confirm that they follow the electronegativity balancing and charge neutrality rules of crystals. We eliminate any stoichiometries currently included in the OQMD database because the goal of this study is to create novel crystals not found in existing material databases and call this step as the OQMD filtration. Since the sites were rounded off before training, some crystals in the dataset lost their symmetry. As a result, we find an inherent lack of symmetry in some generated compounds. The symmetry finder in pymatgen was used to refine the structures to get the nearest symmetric compounds up to 1 angstrom and 30-degree angle

changes. After this process of structure refining, any compounds with two atoms separated by less than 1 angstrom were discarded.

### 2.6. DFT Calculations

DFT calculations were performed to verify the discovered novel crystalline materials. We chose 21 structures at random and optimized them, then calculated their formation energy per atom. The calculations were run on Quantum Espresso 6.7 software using the SSSP-Efficiency pseudopotential library. All atoms were fully relaxed until the force on every atom converged to below 0.001 eV/Å. A 4 × 4 × 4 automatic k-point mesh was used for the calculations. The force was converged within $10^{-3}$ Ry/Bohr, while the convergence threshold for self-consistency was taken to be $10^{-6}$ Ry.

## 3. Results and Discussion

### 3.1. Formation Energy Model and Crystal Structure Representation

Table 1 shows the mean absolute errors of the formation energy models for each representation and their comparison with CGCNN. Rep1 has comparably high MAE in comparison to other representations, which is understandable as it doesn't contain any information about the elemental properties. Notably, the two representations (Rep2 and Rep3) consisting of elemental properties have comparable errors with respect to CGCNN, giving us high confidence in their modeling capabilities. Rep3 performs marginally better than Rep2, but it has a significantly higher number of properties. Discretization of the large number of columns for the generation model would increase the dimensionality of the problem, effectively increasing the computational complexity and simultaneously making it harder for the VAE to correctly generate the samples. Rep2 is one-hot encoded, which makes it more suitable for generative modelling purposes. The decision on the

representation used for the generative model is based on the trade-off between the accuracy of the formation energy model and the dimensionality of the system. Using a higher-dimensional representation might give better results, but would also make it harder to train our generative model. Therefore, we choose Rep2 for training the CVAE.

**Table 1.** *Performance of the formation energy models with different crystal structure representations.*

| Representation | Model | MAE (eV/atom) |
|---|---|---|
| Crystal graphs | CGCNN | 0.026 |
| Rep1 | FNN | 0.045 |
| Rep2 | FNN | 0.028 |
| Rep3 | FNN | 0.026 |

### 3.2. Latent Space Properties

As mentioned earlier, latent space created by the encoder should be a well regularized continuous space so that a point sampled out of the latent space results in a well-defined crystal structure after decoding. Figure 4a shows the kernel density estimate (KDE) plots for all the 100 latent space dimensions. KDE plots are essentially probability density plots, which give an idea about the distribution of a particular variable. It can be observed that the majority of the latent space dimensions are normally distributed, which indicates that the latent space is well regularized and it can be used reliably to sample points to generate new crystal structures.

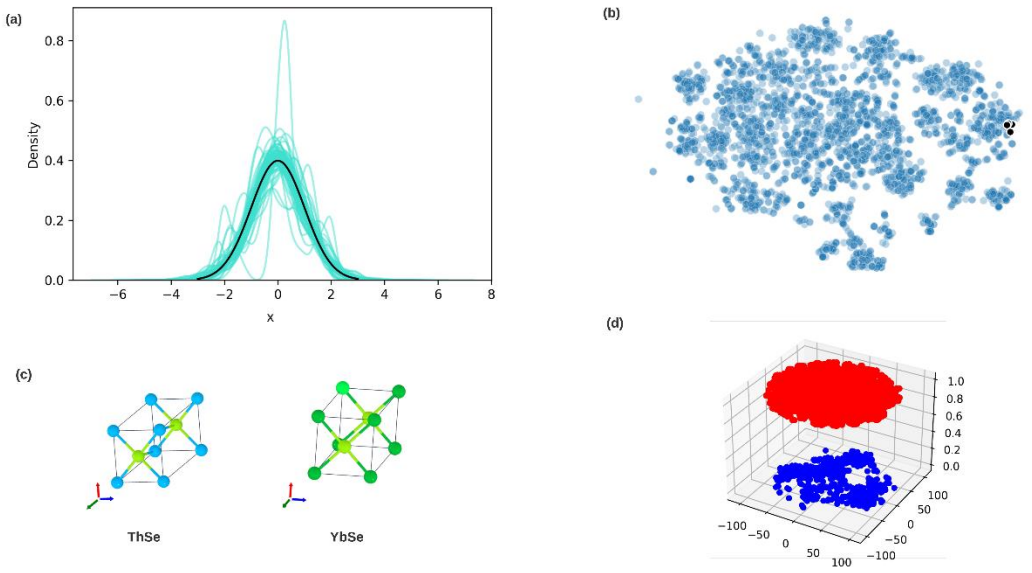


**Figure 4.** *Plots demonstrating the quality of the latent space. (a) KDE plots for 100 latent space dimensions, Black plot is standard normal for reference, all others are latent dimensions. (b) t-SNE encoding of 3000 crystals of the VAE latent space. Black scatter points were sampled out of the latent space to show the continuity of the latent space. Crystals with similar chemical compositions are grouped together. (c) crystal structure of ThSe and YbSe decoded from the scatter points in b (d) 3-dimensional representation of the latent space constructed using CVAE to show the grouping of stable (red) and unstable (blue) crystal structures.*

Figure 4b shows the t-SNE encoding of the 3000 crystals of the VAE latent space. The scatter points marked in black were chosen randomly to demonstrate the continuity of the latent space. After decoding these points from the latent space, it resulted into the crystal structures of following carbides of heavy metals, HfSe, NbSe, PaSe, TaSe, ThSe, TiSe, VSe, ZrSe, PuSe and SmSe. All the points belong to the same crystal structure and are carbides of 3$^{rd}$, 4$^{th}$ and 5$^{th}$ group metals.

Close clustering of crystal structures of similar compounds demonstrates the continuity of the latent space. For all the crystals in the structure, only the elements change in the structure shown above. The crystals are grouped in the latent space according to chemical composition as well as structure. The similarity in structure can be seen in the structures visualized in fig. 4c.

For the reverse engineering of materials with specific properties, the generative models should also be able to cluster compounds based on the target property, i.e. we should be able to generate materials on the basis of formation energy in the present case. To demonstrate the clustering of latent space according to the formation energy, principal component analysis was used to reduce the dimensions of the latent space to 2. To preserve the essence of the latent space in a CVAE, we include a third dimension, information on the stability of our crystal (fig 4d). The clustering of the latent space based on the formation energy can be clearly observed in fig. 4d. The points marked in red are the encodings of stable compounds, whereas the blue markers are the unstable encodings.

### 3.3. CVAE model

Reconstruction accuracy of the CVAE model was calculated to assess its performance. A prediction was counted as correct, if it corresponds to the same stoichiometry and the unit cell. E.g. HCl should be predicted as HCl. $H_2Cl_1$, $HCl_2$, and BaCl are all wrong. As the basis of the unit cell is one hot encoded, a perfect prediction corresponds to predicting the correct bin for each element's position along all the three axes.

The VAE model trained on Rep2 achieved 98.57% accuracy in stoichiometry prediction, 95.57% accuracy of the unit cell and 94.3% accuracy for the prediction of both on the validation dataset. The corresponding numbers were 98.69% accuracy in stoichiometry prediction, 96.59% accuracy of the sites and 95.3% accuracy for the prediction of both on the training dataset. However,

considerable error was observed in the prediction of the lattice parameters of the decoded crystals (table 2), suggesting a need for the geometry optimization.

**Table 2.** *Mean absolute errors in the lattice cell parameters for the training and validation datasets predicted by the CVAE model.*

| | Training MAE | Validation MAE |
|---|---|---|
| a (Å) | 0.38 (8.23%) | 0.38 (8.39%) |
| b (Å) | 0.39 (8.47%) | 0.39 (8.56%) |
| c (Å) | 0.40 (8.20%) | 0.40 (8.37%) |
| α (°) | 11.67 (15.80%) | 11.71 (15.83%) |
| β (°) | 10.73 (14.24%) | 10.91 (14.45%) |
| γ (°) | 10.13 (14.77%) | 10.61 (15.65%) |

It was found that the model rediscovered around 21% stable and 27 % unstable materials present in the training and validation dataset. This rediscovery is encouraging and further verification of the validity of our model. The mean absolute percentage errors for the lattice constants of these rediscovered crystals was found to be similar to the training and validation dataset MAEs, i.e.~15% error in the lattice constants and ~20% MAE for the unit cells angles.

### 3.4. Crystal Structure Generation

So far it has been shown that the latent space is well regularized to sample points from it and CVAE decoder has reasonable accuracy to decode the latent space points into a crystal structure

with a reasonable accuracy. SLERP method was used to sample points from the latent space to generate new crystal structures. Formation energy was used as an input to generate stable or unstable compounds. Points were sampled from the stable part of the latent space (fig 4d) to generate stable crystal structures and vice versa for the unstable crystal structures. Table 3 shows the statistics of the crystal structure generation. Total 11000 and 2250 points were sampled out of the latent space to generate the stable and unstable crystal structures, respectively. The number of runs were chosen according to the ratio of stable and unstable materials in the original dataset.

After decoding the points into crystal structures using the CVAE decoder, they were passed through the SMACT filter to check for the charge neutrality and electronegativity conditions. 18% of the generated stable and 14% of the unstable ones passed through the SMACT filter successfully. This success rate in generating the new materials is significantly higher than the ones reported in previous works. 21% of the correctly generated stable and 38% of the unstable crystal structures were found to be rediscovered, i.e. they already existed in the OQMD database. Correctly rediscovered compounds reconfirm the ability of the model to generate new crystal structures with target formation energy. Rest of the generated materials are novel materials that don't exist in the literature.

**Table 3.** *Statistics of the crystal structures sampled out of the latent space with the targeted formation energy (stable or unstable).*

| | Stable | Unstable |
|---|---|---|
| Runs | 11000 | 2250 |
| Post SMACT filtering | 1973 (18%) | 322 (14%) |

| | | |
|---|---|---|
| Rediscovered | 423 (3.8%) | 89 (3.9%) |
| Out Of Dataset | 1550 (14%) | 233 (10.3%) |

To calculate the probability distribution of the elements, first the stable and unstable materials were sampled from the respective zones of the latent space. Thereafter, the neural network formation energy model was used to predict the formation energy of the valid crystal structures after filtering. Figure 5 shows the frequency based elemental distribution of the OQMD dataset, generated stable and unstable materials for two different threshold energy values. Crystal structures with formation energy within ±0.4 eV from the threshold energy value were used to create the frequency based probability maps. It can be observed that the elemental distribution of the generated stable materials resembles the distribution of the OQMD dataset as the OQMD dataset is heavily dominated by the stable inorganic materials. Relative probability of the elements is also in good agreement, e.g. H, Li, Na, transition metals appear quite prominently in both the probability plots. Commonly found metals (e.g. Al, Ga, In), semiconductors (Si, Ge) also feature significantly in the stable crystal structures. Inert elements (He, Ar etc.) don't appear at all in both the probability plots as they don't form any solid crystals. Absence of some of the elements in the unstable materials is clearly pronounced, e.g. H, Li. Na, which are more frequently found in the stable inorganic materials do not appear in the unstable materials. These intuitive findings give us confidence in the crystal structures generated through the process.

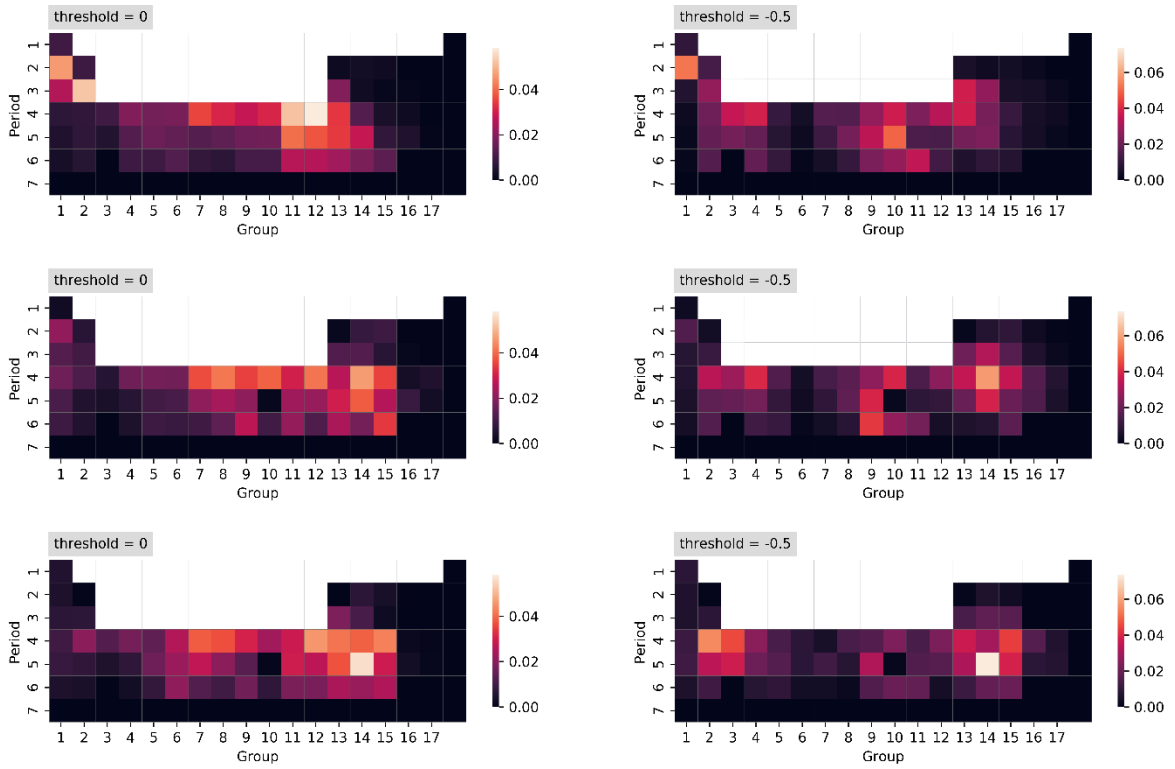


**Figure 5.** *Probability distribution of the periodic table elements in (a) OQMD database, (b) generated stable crystals, and (c) generated unstable crystals, calculated based on the frequency with which the elements appear in the crystal structures within a window of 0.4 units around a chosen threshold formation energy.*

Figure 6 shows the formation energy distribution for the generated stable, unstable and OQMD datasets. It is clear that the distribution for the unstable compounds is very sharply centered around zero (fig. 6b), and it also has a tail extending farther to the right than others, while the distribution for the stable generated compounds (fig. 6a) extends much farther to the left. Shapes of the distributions of the stable and original datasets are also quite similar (fig 6a,c). Primary peak is observed at the same formation energy in both the distributions.

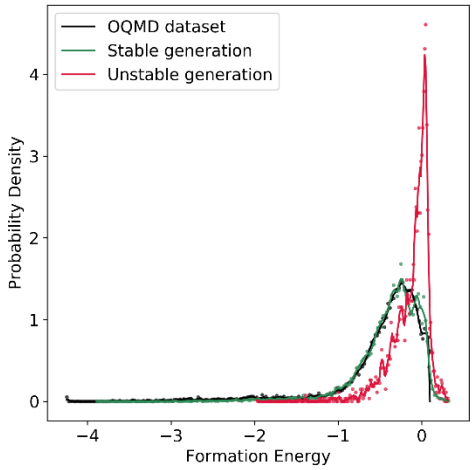


**Figure 6.** *The distribution of formation energies Showing how the CVAE can generate new structures with high probability of achieving target property, even in a highly skewed dataset.*

### 3.5. Validation using DFT Calculations

12 stable and 9 unstable predicted structures were randomly chosen to optimize using DFT calculations. We then define a specific set of rules to determine whether a generated structure is validated as correct. Only if the change in fractional coordinates of each site is less than 1% is a structure regarded accurately predicted. The structure should be accurately predicted, and the formation energy should be negative for a correct stable structure. When the formation energy of a structure is positive and the structure is properly predicted, it is said to be unstable.

Overall 4/12 of the predicted stable structures were found to be stable and 3/9 predicted unstable structures were found to be unstable giving a total accuracy of 33% in searching for novel materials. This result is extremely promising as we are able to find correct novel structures and stoichiometries with high accuracy. Our approach improves upon previous studies that have shown drastic changes in the atomic coordinates after relaxation. However, we note that the predicted structures are still not the lowest energy structures, and some relaxation of lattice parameters is

required. This also means that some structures which were generated as unstable might go to form structures with stable formation energy after relaxation. Figure 7 shows the crystal structures of some of the predicted materials before and after relaxation.

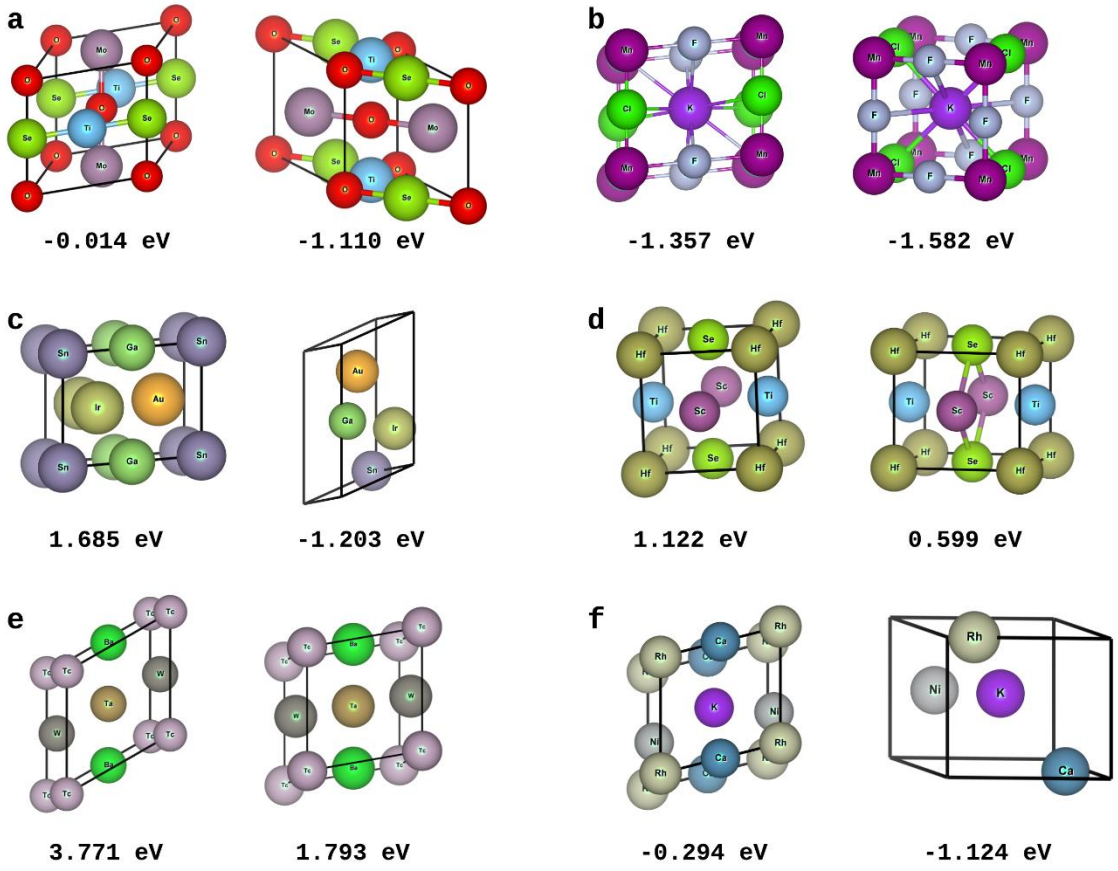


**Figure 7.** *Predicted crystal structures before and after optimization. a,b) Correctly predicted stable materials $TiMoSeO_2$ and $KMnClF_2$. c) an incorrectly predicted stable structure GaSnIrAu. d,e) correctly predicted unstable structures HfScTiSe and BaTaTcW. f) an incorrect unstable structure KCaNiRh.*

## 4. Conclusion

In this paper, we have introduced a scalable and generalizable 2D matrix representation for the crystal structure of the inorganic materials. The representation contains the crystallographic information of the structure as well as the properties of the constituent elements and has no limitation in terms of the type of the crystal structure or the stoichiometry. 2D matrix representation was used to train a conditional variational autoencoder model, which was then used for the targeted generation of the stable or unstable crystal structures based on the value of the formation energy. The CVAE model trained on the OQMD database was able to generate the novel crystal structures, which are not present in the entire database. The generated crystal structures showed a very high accuracy in the prediction of crystallographic sites and the stoichiometry, however the lattice constant prediction needs further improvement. DFT calculations were done to calculate the formation energy of few discovered structures, which showed that the model was able to generate 33% compounds with the correct target property. This study also indicates the possibility of property-driven generation of crystal structures using generative models, which has been demonstrated here for the formation energy. We hypothesize that this can be done for other useful properties, provided the representations are able to model the properties sufficiently accurately.

**Acknowledgements:** Authors would like to thank the funding support from Science and Engineering Research Board grant SRG/2019/000644 and Department of Science and Technology grant DST/TMD/IC-MAP/2K20/03 (C). The HPC facility provided by the institute computer center are also acknowledged.

***References***


(1) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Computational Materials Science* **2012**, *58*, 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005.

(2) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput Mater* **2015**, *1* (1), 1–15. https://doi.org/10.1038/npjcompumats.2015.10.

(3) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323.

(4) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(5) Lu, S.; Zhou, Q.; Chen, X.; Song, Z.; Wang, J. Inverse Design with Deep Generative Models: Next Step in Materials Discovery. *National Science Review* **2022**, *9* (8), nwac111. https://doi.org/10.1093/nsr/nwac111.

(6) Jabbar, R.; Jabbar, R.; Kamoun, S. Recent Progress in Generative Adversarial Networks Applied to Inversely Designing Inorganic Materials: A Brief Review. *Computational Materials Science* **2022**, *213*, 111612. https://doi.org/10.1016/j.commatsci.2022.111612.

(7) Chen, L.; Zhang, W.; Nie, Z.; Li, S.; Pan, F. Generative Models for Inverse Design of Inorganic Solid Materials. *Journal of Materials Informatics* **2021**, *1* (1), 4. https://doi.org/10.20517/jmi.2021.07.

(8) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1* (5), 1370–1384. https://doi.org/10.1016/j.matt.2019.08.017.

(9) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* **2020**, *6* (8), 1412–1420. https://doi.org/10.1021/acscentsci.0c00426.

(10) Pathak, Y.; Singh Juneja, K.; Varma, G.; Ehara, M.; Deva Priyakumar, U. Deep Learning Enabled Inorganic Material Generator. *Physical Chemistry Chemical Physics* **2020**, *22* (46), 26935–26943. https://doi.org/10.1039/D0CP03508D.

(11) Türk, H.; Landini, E.; Kunkel, C.; Margraf, J. T.; Reuter, K. Assessing Deep Generative Models in Chemical Composition Space. *Chem. Mater.* **2022**, *34* (21), 9455–9467. https://doi.org/10.1021/acs.chemmater.2c01860.

(12) Long, T.; Zhang, Y.; Fortunato, N. M.; Shen, C.; Dai, M.; Zhang, H. Inverse Design of Crystal Structures for Multicomponent Systems. *Acta Materialia* **2022**, *231*, 117898. https://doi.org/10.1016/j.actamat.2022.117898.

(13) Court, C. J.; Yildirim, B.; Jain, A.; Cole, J. M. 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning. *J. Chem. Inf. Model.* **2020**, *60* (10), 4518–4535. https://doi.org/10.1021/acs.jcim.0c00464.

(14) Ren, Z.; Tian, S. I. P.; Noh, J.; Oviedo, F.; Xing, G.; Li, J.; Liang, Q.; Zhu, R.; Aberle, A. G.; Sun, S.; Wang, X.; Liu, Y.; Li, Q.; Jayavelu, S.; Hippalgaonkar, K.; Jung, Y.;

Buonassisi, T. An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties. *Matter* **2022**, *5* (1), 314–335. https://doi.org/10.1016/j.matt.2021.11.032.

(15) Noh, J.; Gu, G. H.; Kim, S.; Jung, Y. Machine-Enabled Inverse Design of Inorganic Solid Materials: Promises and Challenges. *Chem. Sci.* **2020**, *11* (19), 4871–4881. https://doi.org/10.1039/D0SC00594K.

(16) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301.

(17) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput Mater* **2016**, *2* (1), 1–7. https://doi.org/10.1038/npjcompumats.2016.28.

(18) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C. Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Comput Mater* **2022**, *8* (1), 1–26. https://doi.org/10.1038/s41524-022-00734-6.

(19) Shoemake, K. Animating Rotation with Quaternion Curves. *SIGGRAPH Comput. Graph.* **1985**, *19* (3), 245–254. https://doi.org/10.1145/325165.325242.

(20) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Skelton, J. M.; Morita, K.; Walsh, A. SMACT: Semiconducting Materials by Analogy and Chemical Theory. *Journal of Open Source Software* **2019**, *4* (38), 1361. https://doi.org/10.21105/joss.01361.

*Table of Content Graphic*