Computational Materials Science 194 (2021) 110436

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Crystal structure prediction in a continuous representative space

In-Ho Lee [a], K.J. Chang [b,*]

[a] Korea Research Institute of Standards and Science, Daejeon 34113, South Korea
[b] Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

ARTICLE INFO

Keywords:
Crystal structure
Radial distribution function
Variational autoencoder

ABSTRACT

Here we report a method of finding multiple crystal structures similar to the known crystal structures of materials on database through machine learning. The radial distribution function is used to represent the general characteristics of the known crystal structures, and then the variational autoencoder is employed to generate a set of representative crystal replicas defined in a two-dimensional optimal continuous space. For given chemical compositions and crystal volume, we generate random crystal structures using constraints for crystal symmetry and atomic positions and directly compare their radial distribution functions with those of the known and/or replicated crystals. For selected crystal structures, energy minimization is subsequently performed through first-principles electronic structure calculations. This approach enables us to predict a set of new low-energy crystal structures using only the information on the radial distribution functions of the known structures.

1. Introduction

In recent years, the design and discovery of new materials have attracted a lot of attention, but they remain a challenge in materials science [1]. Finding stable crystal structures with good functionality is a major concern in material design. First of all, since a large number of crystal structures are possible, the space to explore is immensely large. In addition, there is no guarantee that the proposed crystals can be synthesized experimentally. Nevertheless, material design has made significant progress in recent years compared to the fact that in the past it has been largely relied on the intuition of human experts. One of the recently proposed material design methods is the simultaneous use of first-principles electronic structure calculations and global optimization, in which the material properties can be described from first-principles [2–10]. *Ab initio* random structure searching and high-throughput calculations have also shown productivity [11–13]. The active use of global optimization methods allows us to more directly optimize the desired material properties. The target objectives have been demonstrated for Si and C allotropes with direct band gaps [14–18], semiconductor alloys [19], P allotropes with high mobility [20], topological materials [21–24], and high-pressure superconducting phases [25–28].

Another approach is based on a database of crystal structures that are known experimentally and/or theoretically [29–33]. Hereinafter, the term database refers to experimental or computational data on crystal structures. Machine learning is a kind of artificial intelligence based on the idea that machines can learn and make predictions through big data, from which key information can be uniquely extracted. The development and application of machine learning has soared over the last decade and has been markedly successful in many areas [34–36]. In particular, neural networks are efficient to solve challenging problems, such as natural language processing, image recognition, and translation. With the help of a unique data space, machine learning has succeeded in searching for similar but new molecules. Sampling from the latent space has been used to predict new molecules with the desired properties [29,32]. The predicted molecule or crystal structure and its characteristics can be examined in more detail through first-principles calculations. Recently, some molecules predicted by a special type of variational autoencoders (VAEs) [37,38] have been experimentally validated [39].

There is a growing interest in the use of deep neural networks in materials research to answer questions such as designing new molecules and crystals and understanding their electronic properties [40,41]. In condensed matters, the parameter space, determined by composition, atomic position, and cell volume, can increase enormously, similar to big data sets in image or industrial data analysis. Exploring such complex systems is a difficult task in traditional methods based on human intuition. Given this, deep neural networks are a new application and very powerful in predicting new functional materials [29,31–33].

In this work, we propose a method of exploring unknown crystal structures using the generative VAE method [37,38]. We represent

* Corresponding author.
E-mail address: kjchang@kaist.ac.kr (K.J. Chang).

https://doi.org/10.1016/j.commatsci.2021.110436
Received 13 November 2020; Received in revised form 10 March 2021; Accepted 10 March 2021
Available online 10 April 2021

crystal structures in terms of radial distribution function (RDF), which is usually determined by calculating the distances for all pairs of constituent atoms and binning them into the histogram. The VAE method is used for the first time to learn the RDF characteristics of known crystal structures. In order to find new crystal structures other than the known crystals in discrete forms, we generate various RDF replicas through the VAE prediction process. Thus, unknown crystal structures can be explored in a more continuous space provided by VAE. After RDF replicas are prepared, we independently generate random crystal structures satisfying the symmetry of space groups and then select low-energy crystal structures with the RDFs similar to the known database.

This paper is organized as follows: the details of RDF and Pearson's distance for characterizing an arbitrary crystal structure, the VAE method utilized to create RDF replicas, and generation of random crystal structures using crystal symmetry and Wyckoff positions are given in Section 2. In Section 3, VAE is utilized to represent the RDF characteristics of Si and SnSe crystal structures in the machine learning domain, and the results of searching for new crystal structures are given. The performance of the protocol is monitored and an appropriate selection of machine-learning parameters is made during the process of crystal structure search. Finally, conclusions are given in Section 4 by demonstrating that the present scheme of machine learning can serve as a good starting point to search for unknown crystal structures.

2. Methods

2.1. RDF and Pearson's distance between two crystals

The RDF $n(r)$ represents the mean number of atoms in a shell of width dr at distance r, defined as

$$n(r) = g(r)(\rho 4\pi r^2 dr), \tag{1}$$

where $g(r)$ is the pair correlation function and ρ is the atom density. The RDF describes the bonding characteristics of a crystal regardless of its chemical composition and the choice of unit cells, but lacks information on the distribution of bond angles. For numerical calculations, we used a cut-off radial distance of 8 Å and softened the RDF using a Gaussian broadening scheme with a standard deviation of 0.060 Å, which is an artificial device chosen to ensure the continuity of data.

Using the RDF representation, one can obtain the Pearson's correlation coefficient (r_P) between two crystal structures. This coefficient is just the covariance of two variables divided by the product of their standard deviations, expressed as

$$r_P = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2}}, \tag{2}$$

where $\{x_i\}$ and $\{y_i\}$ are the RDF values of two crystal structures at m grid points in the radial direction and $\bar{x}$ and $\bar{y}$ are their average values. While the r_P values range from -1 to 1, the Pearson's distance, defined as $1-r_P$ for two crystal structures, lies in between 0 and 2. The Pearson's distance has been used in cluster analysis and data detection for communication and storage with unknown gains and offsets.

2.2. RDF replicas

A set of crystal structures that we know of is usually represented in discrete forms [42–44]. In addition, the number of crystal structures in a given data set is limited. In order to find more efficiently various crystal structures with similar RDFs, we need to expand the data space by creating additional RDFs, called RDF replicas, which are similar, but not identical, to those of the database. In doing this, the VAE method is advantageous for representing crystal structures in a more continuous space. To create RDF replicas using the database, we examined the capability of VAE to deal with the complex and discrete nature of the crystal structure space. The key challenge is to extract essential information on the bonding characteristics of known crystal structures.

In the machine learning process, it is important to introduce an appropriate representation of crystal structures. For molecule design, a discrete molecular representation, such as a SMILES string, can be used to capture the key features of molecules [29]. Recently, a different approach has been reported for crystal representation, based on crystal graphs composed of nodes and edges that represent atoms in the unit cell and atom connections, respectively [40]. However, it is still a difficult problem to find a way of representing the structural characteristics of discrete crystal structures for the application of VAE. The Pearson's distance based on the RDF is not an absolute value for finding similarity in structure. However, RDF can be a good physical measure for classifying new crystal structures because it provides information on distances for all pairs of constituent atoms and can also be determined experimentally. Our approach using RDF is useful for creating RDF replicas using a machine learning technique and exploring new crystal structures with specific RDFs.

2.3. Autoencoder and VAE

In artificial neural network, an autoencoder (AE) aims to learn key features for a set of data, usually in the form of dimensionality reduction by training the network [45,35]. Along with the reduction side, the reconstruction side is learned, where AE attempts to produce a representation as close as possible to its original input from the reduced encoding. This AE consists of two parts: an encoder and a decoder that are actively executed with data conversion, such as $\vec{x} \rightarrow \vec{h} \rightarrow \vec{x}'$, where $\vec{x}$ and $\vec{x}'$ represent the input and output RDFs, respectively (Fig. 1). The encoder stage of AE takes the input RDFs for training. The data at the encoder neurons $\vec{h}$ are usually referred to as code, latent variables, or latent representation in a continuous space. Meanwhile, at the decoder stage of AE, data conversion occurs such as $\vec{h} \rightarrow \vec{x}'$. The perfect reconstruction, $\vec{x} = \vec{x}'$, is targeted, but the RDF data may not be fully recovered because of the inherent structural nature of neural networks that restrict information in bottle-neck hidden layers at the intermediate stage.

A generative VAE model is a kind of autoencoder with regular training designed to avoid overfitting, like generative adversarial networks [46]. Thus, the VAE's latent space promises to create new data, for example, RDF replicas. The term 'variational' comes from the close

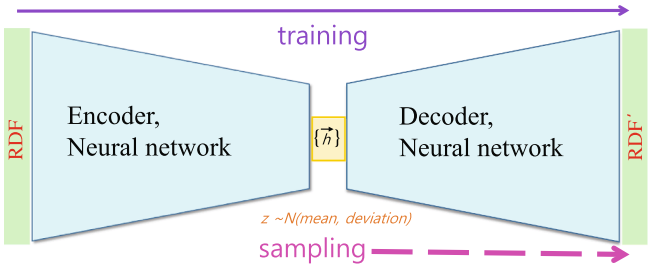

Fig. 1. The autoencoder is composed of the encoder and decoder associated with data conversion, such as $[200, \text{RDF}] \rightarrow 50 \rightarrow 25 \rightarrow 12 \rightarrow \left[2, \{\vec{h}\}\right] \rightarrow 12 \rightarrow 25 \rightarrow 50 \rightarrow [200, \text{RDF}']$. Here RDF and RDF′ correspond to the input and output radial distribution functions, respectively, and $\{\vec{h}\}$ denotes the bottle-neck hidden layer. For a given set of encoder and decoder, the neural network is optimized for about 14,000 trainable parameters, with maintaining maximum information when encoding and minimizing reconstruction error when decoding. Some information is lost during the encoding procedure and cannot be reconstructed when decoding. This is because AE is trained by the VAE learning process that uses a continuous representation in the latent space. The input RDF is encoded as a normal distribution in the latent space.

relationship between regularization and variational inference in statistics. In the VAE model, the distribution of latent variables is assumed in low dimensions, and a variational approach is used to find a normal distribution. With the encoder setting to Gaussian, VAE is built on top of neural networks. The encoder can be represented as a standard neural network by an activation function that maps the original data to the latent space [38]. The decoder then maps the latent space at the bottleneck to the same output as the input. So VAE is also called self-supervised learning. Instead of learning a function, VAE learns parameters in the probability distribution representing the data. New data can be generated by sampling from the latent z-space provided by VAE. The probability distribution of the latent vector better matches that of the training data compared to the standard AE.

During the training process, we adopted two assumptions [37,38]: for a training set $\vec{x}$, data are generated using a directed graphic model $p_\theta(\vec{x}|\vec{h})$, and the encoder learns an approximation $q_\phi(\vec{h}|\vec{x})$ to the posterior distribution $p_\theta(\vec{h}|\vec{x})$, where ϕ and θ denote the parameters of the encoder and decoder, respectively. We minimized the loss function of VAE, expressed as

$$L(\phi,\theta,\vec{x}) = D_{KL}\left[q_\phi(\vec{h}|\vec{x})||p_\theta(\vec{h})\right] - E_{q_\phi(\vec{h}|\vec{x})}\left[logp_\theta(\vec{x}|\vec{h})\right]. \quad (3)$$

The first term, Kullback–Leibler divergence [47], quantifies how much one probability distribution differs from another distribution. The prior distribution of latent variables is usually set to a centered Gaussian $p_\theta(\vec{h})$, which is isotropic in the latent space. In general, this regularization is referred to as the problem of calculating the statistical distance between two statistical objects, such as probability distribution. The second term is given by an expectation value over the probability $p_\theta(\vec{x}|\vec{h})$. This reconstruction term relies on the fact that $\vec{x}$, which is extracted from the encrypted $\vec{h}$, should be well represented through a probability distribution. For a given input $\vec{x}$, the probability should be maximized such as $\vec{x}' = \vec{x}$ by minimizing the loss function, when sampling $\vec{h}$ from the distribution $q_\phi(\vec{h}|\vec{x})$ and then sampling $\vec{x}'$ from the distribution $p_\theta(\vec{x}|\vec{h})$. The reconstruction term makes both encoding and decoding efficient, while the regularization term leads to a regularly organized latent space.

Numerical calculations were performed using a deep learning library, Keras [48], which is an advanced neural networks application programming interface, written in Python. This library can run on top of TensorFlow [49], CNTK [50], or Theano [51] and was developed to allow fast experimentation. We used 'Binary-crossentropy' and 'Adam' optimizer in Keras [48]. The step size, batch size, and other parameters in the optimizer did not significantly alter the self-supervised learning performance in the present experiment. A rectified linear unit was used for the activation function. The initialization of weights was based on the normal distribution function. We also ran a series of tests on changing the number of layers and the dropout probability of dropout layers, and found that the self-supervised learning process was very stable.

The encoder consisting of artificial neural networks reduces the dimension of RDF information from 200 to 50, 25, 12, and finally 2, as illustrated in Fig. 1. Similarly, the decoder was designed with artificial neural networks that control the data flow, increasing the dimension from 2 to 12, 25, 50, and finally 200. Here the input dimension of 200 was chosen as a parameter to discretize the radial distance and varies with the cut-off distance. We used a set of fully connected neural networks. Neurons in each layer are connected to all activations in the previous layers. Thus their activations can be calculated with the matrix–vector multiplication followed by a bias offset. The total number of parameters in the neural networks is around 14,000. It was an easy optimization problem, so there was no difficulty in the training process.

RDF replicas were created after the training was over. The generation of RDF replicas using VAE-trained networks is analogous to evolutionary algorithms, in which new configurations are generated via crossover and mutation. A neural network was used to learn a representation that maps samples from the input space to the latent z-space, where the distribution of latent vectors is approximated by Gaussian. In fact, generative RDF replicas can be obtained from the trained model by feeding the reduced latent variables, z_1 and z_2, to the learned generative model. A variety of RDFs are needed to create multiple crystal structures as close as to the database. Thus, we chose various z-space distributions, based on the VAE, by introducing a controllable confidence interval in the Gaussian form. Using a 90% confidence interval in the Gaussian distribution, defined in the latent space, we were able to obtain various crystal structures with RDFs similar to those of the database. This procedure is based on the capabilities of the VAE method in the latent space [29], such as interpolation between two points and RDF generation from one RDF to another RDF. Typically, new RDFs are made using random points in the latent space. Going one step further, points in the latent space are selected and then used as inputs or queries to generate specific RDFs. We can create a series of points on a linear path between two points. These points can be used to generate a series of RDFs representing the transition between the two RDFs already known. In addition, these points can be kept in the latent space and used in simple vector arithmetic to create new points and then generate new RDFs. This strategy enables intuitive and targeted RDF generation.

2.4. Generation of random crystal structures

Fig. 2 shows the flowchart of the present calculations. After the RDF replicas were prepared, random crystal structures were independently generated and then their RDFs were compared with those of the database and replicas to examine the structural similarity. Random crystal structures with specific crystal symmetry and chemical compositions can be generated using the existing packages such as AMADEUS [10], AIRSS [11], RanSpg [52], and PyXtal [53].

In three-dimensional space, the symmetry of crystals can be described with the space groups. For a given crystal with a specific space group, atomic positions can be classified by Wyckoff positions [54]. For 230 space groups, numerous crystal structures can be created by changing relative atomic positions. We used our own algorithm implemented in AMADEUS [10] to select possible combinations of Wyckoff positions. For given chemical compositions, a specific space group is first assigned and Wyckoff positions are randomly generated. Then, atoms are added to the unit cell according to the site symmetry. This procedure is repeated until all atomic positions satisfy the desired symmetry and chemical compositions. If an atom added to the unit cell is too close to the other existing atom, it is discarded, and a new combination of Wyckoff positions is generated using the minimum distance between atoms permitted by atomic radii for a given material. This condition is used to extract crystal structure and is already well embedded in the RDF function. Crystal structure generation was done using a stochastic approach that allows parallel computations for practical applications. Using a large number of directories, we explored simultaneously random crystal structures with the same input in each directory. We found that parallel efficiency was very high because communication between directories was not used. Whenever a random crystal structure was generated, its RDF was immediately compared with those of the known database and replicas.

Since the input RDF vector can be defined regardless of its chemical composition, our scheme can be applied to binary and ternary systems, etc. In material design using evolutionary algorithms, crossover and mutation operators are widely used and have been efficient in exploring vast crystalline structures. Because these operators often produce similar crystal structures with very close interatomic distances, similar trial solutions are adjusted by performing proper relaxation operations to ensure proper distances between adjacent atoms [6–10]. However, this adjustment is not required in the current approach that uses space

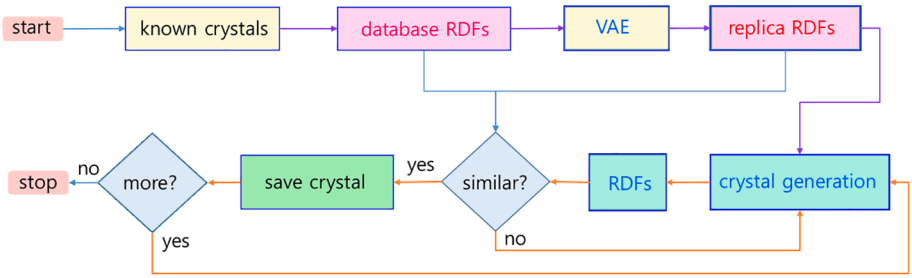

Fig. 2. The flowchart of the crystal structure search based on VAE is shown. VAE is used to generate the RDFs of arbitrary crystal structures similar to those of known crystal structures in a continuously defined space. VAE learns to map RDFs in the latent space and randomly generates new RDFs. Based on the RDFs from the database and VAE replicas, random symmetrical crystal structures are explored through parallel computations, each satisfying one of the 230 space groups in three dimensions. Orange arrows in the flowchart indicate that computations can be parallelized. Two cyanic diamonds represent judgment conditions in the calculation process. Two pink boxes labeled RDFs and replica RDFs make up the RDF set for the known crystal structures of the database and the VAE replicas, respectively. Two light green boxes represent the generation of random crystal structures and their RDFs, respectively.

groups and Wyckoff positions.

2.5. First-principles calculations

Our goal is to find new crystal structures that not only resemble known crystal structures, but also have relatively low energies. In this regard, our combined approach using the VAE method and RDF would be very useful because it filters out similar crystal structures prior to first-principles calculations. To select low-energy crystal structures and investigate their electronic properties, we performed additional first-principles calculations, in which the projector augmented-wave pseudopotentials were used [55], as implemented in the VASP code [56]. We employed the local density approximation (LDA) for the exchange–correlation potential [57] within the density functional theory [58]. The energy minimization for given ionic positions was carried out by calculating the stress tensor [59–61] and Hellmann–Feynman forces [62–64]. Using a $\vec{k}$-point mesh with a grid spacing of $2\pi \times 0.02$ Å^{-1}, the iterative procedure was continued until all the forces were less than 0.01 eV/Å. It should be noted that a high kinetic energy cut-off of 500 eV is required for accurate stress tensor calculations. Here the numerical accuracy of a few kbar was taken for stress calculations.

3. Results and discussion

3.1. Si allotropes

For Si, 36 crystal structures are currently provided on the Materials Project database [44]. Based on the data space, we first created RDF replicas using a confidence interval of 90% in the Gaussian distribution in the latent space. With this variable parameter, we can generate various RDF replicas. The objective of the present VAE is to find a proper projection method that maps RDFs from a high feature space to a low feature space. Representing RDF in a low-dimensional space has an advantage in improving the encoding and decoding performance for crystal classification. In addition, in-depth AEs can exponentially reduce the amount of training data required for RDF learning. We used linearly spaced 15 coordinates for the latent variables, which are transformed through the inverse cumulative distribution function of the Gaussian. Despite multi-dimensions are possible, we chose the latent variables in two-dimensions, z_1 and z_2, each of which ranges up to 15, yielding 225 RDF replicas.

We checked the structural similarity between the database and RDF replicas by calculating Pearson's distances. In the replica samples, the mean Pearson's distance and standard deviation are estimated to be 0.200 and 0.152, respectively, while the corresponding values of the database are 0.344 and 0.161. For the whole samples consisting of the database and replicas, the mean Pearson's distance and standard deviation are 0.283 and 0.156, respectively. This result indicates that the replica RDFs are properly produced with a high degree of similarity.

In Fig. 3, the loss function is plotted as a function of epoch for the

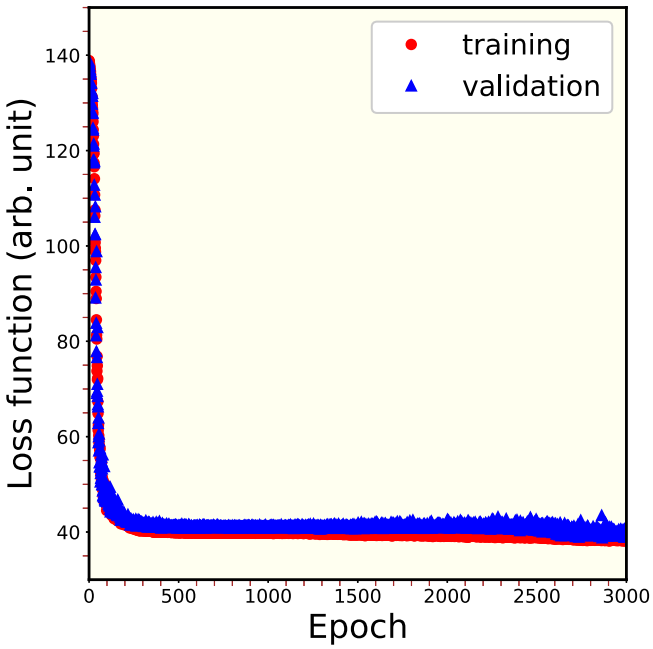

Fig. 3. The loss functions for the training and validation sets are plotted as a function of epoch. We choose 28 and 8 samples from the database of Si allotropes for the training and validation sets, respectively. In the loss function, the reconstruction term makes the encoding–decoding scheme efficient, while the regularization term makes the latent space regular.

training and validation sets, which are about 80% and 20% of the database, respectively. The validation set shows almost the same shape and features as the training set, verifying that the VAE model used well describes the RDFs of the untrained data. In Fig. 4, the energies of the known structures on the Materials Project database [44] are plotted in the two-dimensional latent space reduced by VAE. The energies of various Si allotropes relative to the cubic diamond structure in the database are color-coded. This plot is a good example of multi-dimensional scaling and exhibits an approximately linear relationship between the reduced latent variables, z_1 and z_2. Moreover, it shows the best pair of encoder and decoder that keeps maximum information when encoding and has minimum reconstruction error when decoding.

Next we generated random crystal structures using crystal symmetry and Wyckoff positions for two unit cells containing 8 and 16 Si atoms, as introduced in Section 2.4. We chose the interatomic distances above 2.1 Å and a tolerance of 10% for the crystal volume, *e.g.*, 184 Å^3 for the 8-atom unit cell. The RDFs of random crystal structure were compared with those of the database and replicas. The structural similarity was determined using a criterion of 0.13 for the Pearson's distance, which ensures about 90% similarity. Compared with the RDFs of the known

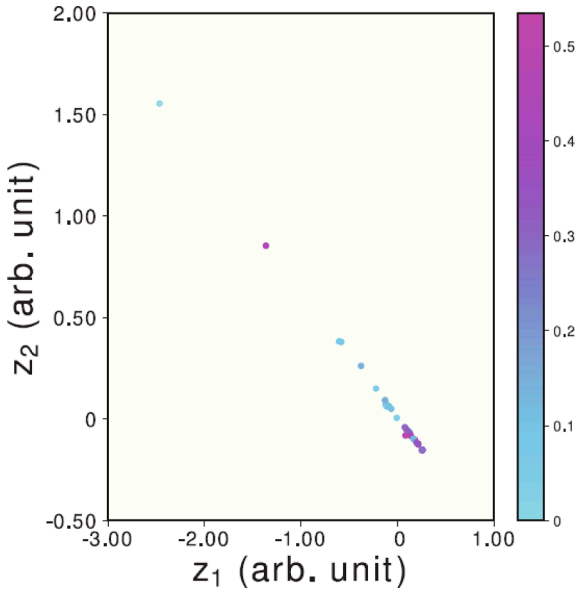

Fig. 4. A two-dimensional representation is used to plot the distribution of 36 known Si allotropes which are usually defined in a multi-dimensional crystal structure space. The relative energies (in unit of eV/atom) of allotropes with respect to cubic diamond Si [44] are color-coded.

database, replicas, and both, the relative proportions of selected crystal structures with similar RDFs but not yet fully relaxed were found to be 22%, 33% and 45%, respectively, after 8 independent runs. This ratio depends on the training set selected in the VAE procedure, but in this way we can see whether any random structure is similar to that in the database or replica samples. In some cases, we found several crystal structures identical to the existing database, for example, cubic diamond $Fd\bar{3}m$ (No. 227), hexagonal diamond $P6_3/mmc$ (No. 194), and monoclinic $C2/m$ (No. 12) phases. For a particular random structure with the space group $P6/mmm$ (No. 191), its RDF is compared in Fig. 5 with that of the known crystal structure with the space group $Im\bar{3}m$ (No. 229), named mp-1072544 in the Materials Project database. The Pearson's distance between the two RDFs is 0.100.

To demonstrate the usefulness of RDF in exploring new crystal structures, we compare the distribution of total energies for three groups of random crystal structures in Fig. 6. In groups I and II, random crystal structures were selected using the RDFs in the database and replicas, respectively, while any RDF was not used in group III. Using the 8-atom cell, we chose 350 crystal structures in each group that meet the Pearson's distance criterion, and then fully optimized them through first-principles calculations. After full relaxations, the average variation in Pearson's distances is 0.2 and the energies are lowered by an average of 0.5 eV/atom. Among 36 known structures in the database, there are 16 allotropes with the energies below 0.2 eV/atom relative to the cubic diamond phase, as shown in Fig. 6. We focus on the energy interval of 0.05–0.20 eV/atom in the low energy region. In this energy segment, we found 14 crystal structures in groups I and II each, whereas 5 in group III. Many low-energy structures are missing in the random sampling approach, but the two methods that utilize the database and replica RDFs complement each other to find low-energy structures. We investigated the novelty of the crystal structure by making a one-to-one correspondence between the crystal structures discovered in groups I and II. In the low energy region, three crystal structures found in group I were also found in group II. For other crystal structures, the mean and standard deviation of the Pearson's distances are estimated to be about 0.2 and 0.1, respectively, indicating that novel crystal structures can be

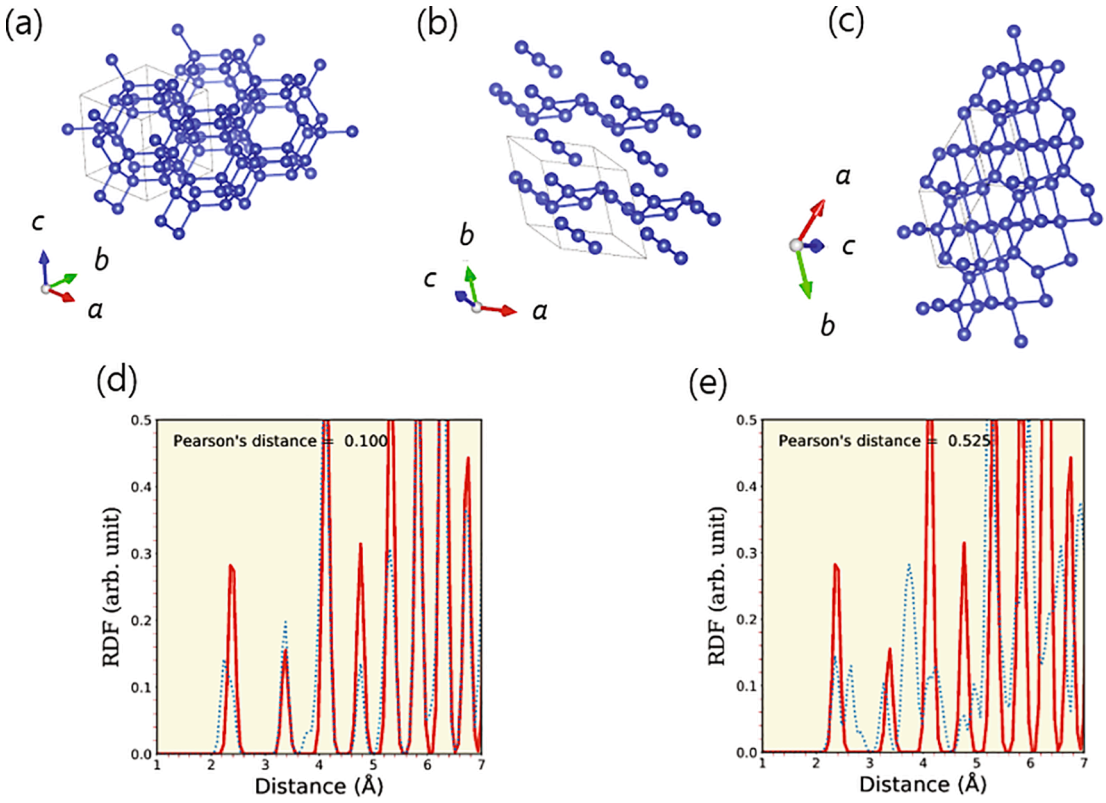

Fig. 5. The atomic structures of (a) one known crystal structure with the space group $Im\bar{3}m$ (No. 229) in the Materials Project database, (b) an arbitrary random crystal with the space group $P6/mmm$ (No. 191), and (c) a fully optimized structure with the space group $Cmmm$ (No. 65). The RDF of $Im\bar{3}m$-Si (red solid line) is compared with those of (d) $P6/mmm$-Si and (e) $Cmmm$-Si (blue dotted line). The Pearson's distances are calculated to be 0.100 and 0.525 for the unrelaxed and fully optimized geometries of the random crystal structure, respectively.

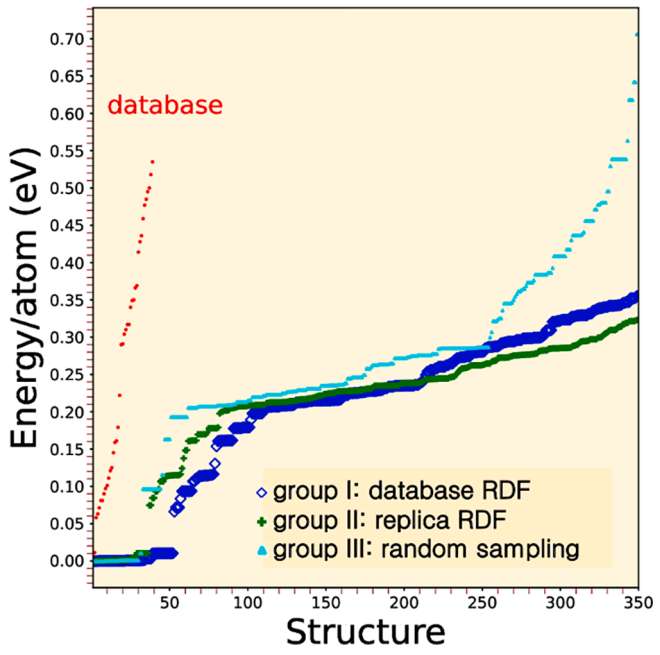

Fig. 6. The energies of 36 known Si crystal structures (red dots) on the Materials Project database [44] are compared with those of crystal structures generated for a supercell containing 8 atoms. In the database, there are 16 low-energy allotropes with a variety of unit cells, such as 0.011 eV (4), 0.058 eV (68), 0.063 eV (46), 0.071 eV (16), 0.081 eV (40), 0.081 eV (46), 0.090 eV (24), 0.097 eV (94), 0.101 eV (8), 0.111 eV (82), 0.121 eV (232), 0.125 eV (164), 0.145 eV (58), 0.159 eV (16), 0.161 eV (24), and 0.179 eV (106), where numbers in parentheses represent the number of atoms per unit cell. The overall energy continuity of the target structures, generated using VAE and subsequently fully optimized through first-principles calculations, matches well that of the database. In group I (blue diamonds), random crystal structures are selected using the RDFs of the known crystal structures in the database, whereas the replica RDFs are used in group II (green crosses). In group III (cyan triangles), crystal structures are randomly generated without using any RDF. In each group, 350 crystal structures are chosen and fully optimized through first-principles calculations.

generated using replica RDFs. We point out that RDF is vulnerable to distinguishing space groups in a crystal structure because the space group can change when the crystal structure is optimized. However, out results show that RDF is useful to produce many random crystal structures on a more continuous scale in the low energy region.

We examined the electronic properties of the selected random crystal structures with energies below about 0.4 eV/atom relative to cubic diamond Si, and found both metallic and semiconducting crystals with various band gaps up to about 1 eV. Among new Si crystal structures, we focus on two specific crystals with the *Cmmm* (No. 65) and $C2/m$ (No. 12) space groups, called *Cmmm*-Si and $C2/m^*$-Si, respectively. Through first-principles calculations, we obtained the *Cmmm*-Si allotrope by fully relaxing the atomic positions and cell shape of the random crystal structure with the space group $P6/mmm$ (Fig. 5), which was first generated using only crystal symmetry and Wyckoff positions and then selected with the criterion of 0.13 for the Pearson's distance. In a fully optimized geometry, the lattice parameters and Wyckoff positions of *Cmmm*-Si are given in Table 1.

In Fig. 5, the RDF of *Cmmm*-Si is compared with that of the known mp-1072544 crystal in the Materials Project database, and the Pearson's distance increases to 0.525, as compared to the value of 0.100 for the unrelaxed structure with the space group $P6/mmm$. It is interesting to note that *Cmmm*-Si is metallic due to the mixing of five- and sixfold coordinated Si atoms. Although *Cmmm*-Si is higher in energy by about 0.35 eV/atom than cubic diamond Si, we found that *Cmmm*-Si satisfies six criteria for the mechanical stability of an orthorhombic structure,

Table 1
Lattice parameters and Wyckoff positions for three Si allotropes, *Cmmm*-Si, $C2/m$-Si, and $C2/m^*$-Si.

Allotrope	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	Wyckoff positions
Cmmm-Si	8.21	6.59	5.20	90.0	90.0	90.0	4 l mm2 (0, 0.50000, 0.75536) 8 m ..2 (0.25000, 0.25000, 0.72747) 4 g 2 mm (0.86311, 0.00024, 0.99999)
$C2/m$-Si	13.73	3.82	6.29	90.0	83.3	90.0	4 i m (0.94460, 0, 0.87347) 4 i m (0.44097, 1.00000, 0.65322) 4 i m (0.78744, 0, 0.05727) 4 i m (0.27303, 1.00000, 0.58547)
$C2/m^*$-Si	3.77	16.56	6.93	90.0	133.7	90.0	4 i m (0.08137, 0, 0.86120) 8 J 1 (0.07598, 0.87869, 0.36283) 4 g 2 (0.00029, 0.29131, 0.00058)

[65–67] maintaining its crystal structure even at ambient pressure.

The atomic structure of $C2/m^*$-Si is very similar to that of a $C2/m$-Si allotrope, named mp-1079297 in the Materials Project database, as shown in Table 1 and Fig. 7. The Pearson's distance between $C2/m^*$-Si and $C2/m$-Si is 0.102 for their fully optimized structures, and all Si atoms in both structures form tetrahedral bonds with the same space group $C2/m$. We found that the monoclinic crystal of $C2/m^*$-Si satisfies twelve criteria for mechanical stability [65–67]. The energy of $C2/m^*$-Si is slightly higher by 40 meV/atom than that of $C2/m$-Si, while its crystal volume is smaller with a density of 2.39 g/cm^3, as compared to the density of 2.28 g/cm^3 in $C2/m$-Si. Both the $C2/m^*$-Si and $C2/m$-Si allotropes have the indirect LDA band gaps of 0.28 and 0.13 eV, respectively, and exhibit very similar RDF characteristics and X-ray diffraction patterns, especially at neighboring distances and small angles, as shown in Fig. 7. Recently, based on microexplosion experiments, two tetragonal ($t32$ and $t32^*$) and two monoclinic ($m32$ and $m32^*$) phases, similar to the BC8 structure, have been proposed on Si surface exposed to ultrashort laser pulses [68]. We found that the $C2/m^*$-Si allotrope is more stable by 3–10 meV/atom than these four phases.

3.2. SnSe allotropes

SnSe is an anisotropic thermoelectric material recently discovered [69,70]. For 8 known structures of SnSe on the Materials Project database [44], we created RDF replicas using the same VAE method and then generated random crystal structures for a supercell containing 12 formula units. We chose the minimal interatomic distances of 2.0, 2.1, and 2.2 Å for the Sn-Sn, Sn-Se, and Se-Se bonds, respectively, and the unit cell volume of 686 Å^3. Despite the small size of the database, many crystal structures were obtained on a somewhat continuous energy scale. Nevertheless, the database is too small for machine learning. To expand the database, we generated crystal structures with relatively low energies (< 0.1 eV per formula unit) through theoretical calculations for

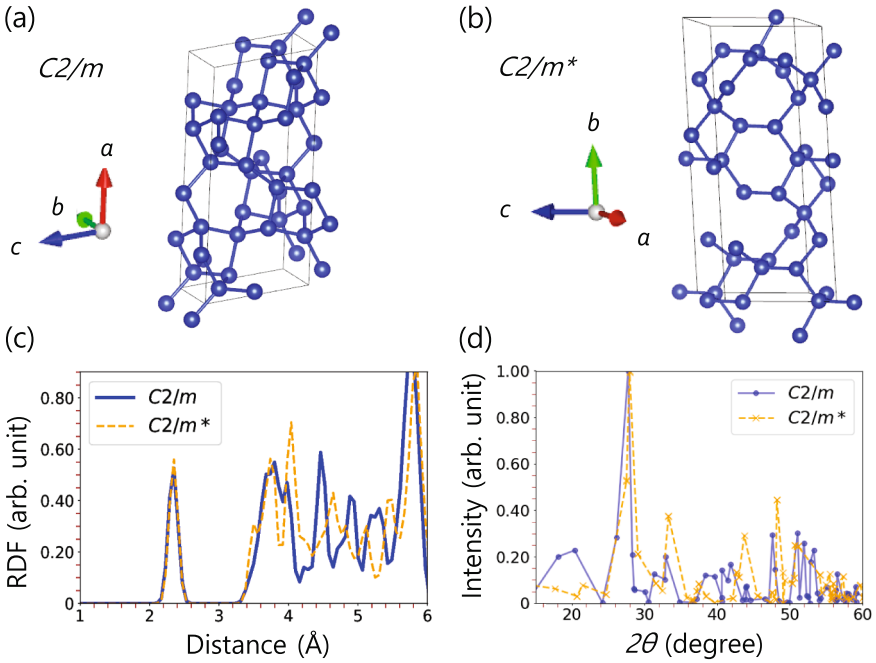

Fig. 7. Atomic structures of (a) $C2/m$-Si and (b) $C2/m^*$-Si. Black solid lines represent the primitive unit cell. The (c) RDFs and (d) X-ray diffraction patterns of $C2/m$-Si and (b) $C2/m^*$-Si are compared.

a unit cell containing 4 formula units using the design code AMADEUS [10]. We selected 50 reference crystal structures and searched for random crystal structures for a supercell containing 8 formula units, as shown in Fig. 8. Similar to the results of Si, we found a variety of crystal structures with nearly continuous energies in a low-energy window below 0.01 eV per formula unit.

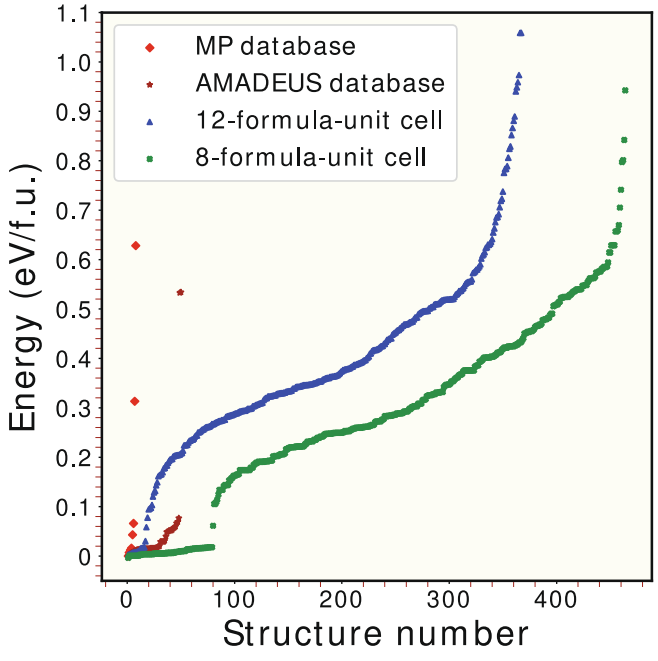

Fig. 8. The energies of 8 known SnSe crystal structures (red dots) on the Materials Project database [44] and 50 low-energy allotropes (dark brown dots) generated using the AMADEUS code [10] are compared with those of randomly generated crystal structures for two different cells containing 8 (green dots) and 12 (blue dots) formula units. The overall energy continuity of the target structures, generated using VAE and subsequently fully optimized through first-principles calculations, matches well that of the database.

4. Conclusion

We have developed a way of finding random crystal structures similar to the known crystal structures on the database using the machine learning technique. The strategy for crystal structure search is as follows: First, the radial distribution function is used as a one-dimensional representation of crystal structures. Second, the Pearson's distance between two radial distribution functions is defined as a measure of similarity between two corresponding crystal structures. Third, the variational autoencoder is used to determine the overall characteristics of the known crystal structures and to generate various replicas with similar RDFs. Finally, for given chemical compositions and cell volume, random crystal structures are independently generated using the crystal symmetry and Wyckoff positions, and then a set of new crystal structures are selected by directly comparing with the RDFs of the database and replicas. Since the RDF is used, this is an indirect approach of searching for arbitrary crystal structures, in which a direct representation of crystal structures is avoided in the continuous space. Our approach can be extended to a conditional material design that generates a new crystal with the structural characteristics predesignated from the known crystals. If the database provides a limited set of allotropes, our VAE method has difficulty in finding new crystal structures that are completely different from the data set. Nevertheless, our approach has been shown to be very efficient for exploring unknown low-energy crystal structures using only information on the interatomic distance and Pearson's distance based on radial distribution function.

Data availability

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

CRediT authorship contribution statement

In-Ho Lee: Methodology, Investigation, Data curation, Visualization, Writing - original draft. **K.J. Chang:** Supervision, Funding acquisition, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Korea Institute for Advanced Study (KIAS Center for Advanced Computation) for providing computing resources. KJC was supported by Creative Materials Discovery Program through the NRF funded by the Ministry of Science and ICT (2018M3D1A1058754).

References

[1] A. Franceschetti, A. Zunger, Nature (London) 402 (1999) 60–63.
[2] D.J. Wales, J.P.K. Doye, J. Phys. Chem. A 101 (1997) 5111–5116.
[3] R. Martoák, A. Laio, M. Parrinello, Phys. Rev. Lett. 90 (2003), 075503.
[4] J. Lee, I.-H. Lee, J. Lee, Phys. Rev. Lett. 91 (2003), 080201.
[5] S. Goedecker, J. Chem. Phys. 120 (2004) 9911–9917.
[6] A.R. Oganov, C.W. Glass, J. Chem. Phys. 124 (2006), 244704.
[7] Y. Wang, J. Lv, L. Zhu, Y. Ma, Phys. Rev. B 82 (2010), 094116.
[8] A.R. Oganov (Ed.), Modern Methods of Crystal Structure Prediction, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2010.
[9] D.C. Lonie, E. Zurek, Comput. Phys. Commun. 182 (2011) 372–387.
[10] I.-H. Lee, Y.J. Oh, S. Kim, J. Lee, K.J. Chang, Comput. Phys. Commun. 203 (2016) 110–121.
[11] C.J. Pickard, R.J. Needs, J. Phys.: Cond. Mat. 23 (2011), 053201.
[12] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, Comput. Mater. Sci. 58 (2012) 218–226.
[13] K. Yang, W. Setyawan, S. Wang, M.B. Nardelli, S. Curtarolo, Nat. Mater. 11 (2012) 614–619.
[14] S. Botti, J.A. Flores-Livas, M. Amsler, S. Goedecker, M.A.L. Marques, Phys. Rev. B 86 (2012), 121204.
[15] H.J. Xiang, B. Huang, E. Kan, S.-H. Wei, X.-G. Gong, Phys. Rev. Lett. 110 (2013), 118702.
[16] I.-H. Lee, J. Lee, Y.J. Oh, S. Kim, K.J. Chang, Phys. Rev. B 90 (2014), 115209.
[17] Y.J. Oh, I.-H. Lee, S. Kim, J. Lee, K.J. Chang, Sci. Rep. 5 (2015) 18086.
[18] Y.J. Oh, S. Kim, I.-H. Lee, J. Lee, K.J. Chang, Phys. Rev. B 93 (2016), 085201.
[19] S.V. Dudiy, A. Zunger, Phys. Rev. Lett. 97 (2006), 046401.
[20] W.H. Han, S. Kim, I.-H. Lee, K.J. Chang, J. Phys. Chem. Lett. 8 (2017) 4627–4632.
[21] S. Kim, W.H. Han, I.-H. Lee, K.J. Chang, Sci. Rep. 7 (2017) 7279.
[22] H.-J. Sung, S. Kim, I.-H. Lee, K.J. Chang, NPG Asia Materials 9 (2017), e361.
[23] G.-M. Kim, H.-J. Sung, W.H. Han, I.-H. Lee, K.J. Chang, J. Phys. Chem. C 123 (2019) 1839–1845.
[24] M. Kim, J. Kim, I.-H. Lee, W.H. Han, Y.C. Park, W.Y. Kim, B. Kim, J. Suh, Nanoscale 11 (2019) 5171–5179.
[25] C.J. Pickard, R.J. Needs, Phys. Rev. Lett. 97 (2006), 045504.
[26] G. Gao, A.R. Oganov, P. Li, Z. Li, H. Wang, T. Cui, Y. Ma, A. Bergara, A.O. Lyakhov, T. Iitaka, G. Zou, Proc. Nat. Aca. Sci. (USA) 107 (2010) 1317–1320.
[27] G. Gao, A.R. Oganov, A. Bergara, M. Martinez-Canales, T. Cui, T. Iitaka, Y. Ma, G. Zou, Phys. Rev. Lett. 101 (2008), 107002.
[28] H.-J. Sung, W.H. Han, I.-H. Lee, K.J. Chang, Phys. Rev. Lett. 120 (2018), 157001.
[29] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, ACS Cent. Sci. 4 (2018) 268–276.
[30] J.P. Janet, L. Chan, H.J. Kulik, J. Phys. Chem. Lett. 9 (2018) 1064–1071.
[31] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, G. Klambauer, J. Chem. Inf. Mod. 58 (2018) 1736–1741.
[32] S. Kang, K. Cho, J. Chem. Inf. Mod. 59 (2019) 43–52.
[33] J.G. Freeze, H.R. Kelly, V.S. Batista, Chem. Rev. 119 (2019) 6595–6612.
[34] Y. LeCun, Y. Bengio, G. Hinton, Nature (London) 521 (2015) 436–444.
[35] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, USA, 2016.
[36] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Nature (London) 529 (2016) 484–489.
[37] D.P. Kingma, M. Welling, Proc. 2nd ICLR (2014) arXiv:1312.6114.
[38] C. Doersch, arXiv:1606.05908.
[39] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, M.S. Veselov, V.A. Aladinskiy, A. V. Aladinskaya, V.A. Terentiev, D.A. Polykovskiy, M.D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R.R. Shayakhmetov, A. Zhebrak, L.I. Minaeva, B. A. Zagribelnyy, L.H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, Nat. Biotechnol. 37 (2019) 1038–1040.
[40] T. Xie, J.C. Grossman, Phys. Rev. Lett. 120 (2018), 145301.
[41] J. Noh, J. Kim, H.S. Stein, B. Sanchez-Lengeling, J.M. Gregoire, A. Aspuru-Guzik, Y. Jung, Matter 1 (2019) 1370–1384.
[42] R. Allmann, R. Hinek, Acta Cryst. A 63 (2007) 412–417.
[43] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Comput. Mater. 1 (2015) 15010.
[44] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, APL Mater. 1 (2013), 011002 https://www.materialsproject.org.
[45] M.A. Kramer, AIChE J. 37 (1991) 233–243.
[46] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680.
[47] S. Kullback, R.A. Leibler, Ann. Math. Stat. 22 (1951) 79–86.
[48] F. Chollet et al. 2015 Keras, https://github.com/keras-team/keras.
[49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, R. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from (2015) tensorflow.org.
[50] A. Linn (25 October 2016). "Microsoft releases beta of Microsoft Cognitive Toolkit for deep learning advances. microsoft.com. Microsoft. Retrieved 30 January 2017. "Title: Microsoft releases beta of Microsoft Cognitive Toolkit.
[51] Theano Development Team 2016 Theano: A Python framework for fast computation of mathematical expressions, arXiv:1605.02688.
[52] P. Avery, E. Zurek, Comput. Phys. Commun. 213 (2017) 208–216.
[53] S. Fredericks, D. Sayre, Q. Zhu, Comput. Phys. Commun. 261 (2021), 107810.
[54] J.R. Hook, H.E. Hall, Solid State Physics, Manchester Physics Series, (2nd ed.)., John Wiley & Sons, 2010.
[55] G. Kresse, D. Joubert, Phys. Rev. B 59 (1999) 1758–1775.
[56] G. Kresse, J. Furthmüller, J. Comput. Mater. Sci. 6 (1996) 15–50.
[57] D.M. Ceperley, B.J. Alder, Phys. Rev. Lett. 45 (1980) 566–569.
[58] P. Hohenberg, W. Kohn, Phys. Rev. 136 (1964) B864–B871.
[59] O.H. Nielsen, R.M. Martin, Phys. Rev. Lett. 50 (1983) 697–700.
[60] I.-H. Lee, S.-G. Lee, K.J. Chang, Phys. Rev. B 51 (1995) 14697–14700.
[61] D.R. Hamann, X. Wu, K.M. Rabe, D. Vanderbilt, Phys. Rev. B 71 (2005), 035117.
[62] H. Hellmann, Einführung in die Quantenchemie, Franz Deuticke, Leipzig, 1937, p. 285.
[63] R.P. Feynman, Phys. Rev. 56 (1939) 340–343.
[64] J. Ihm, A. Zunger, M.L. Cohen, J. Phys. C: Solid State Phys. 12 (1979) 4409-4422; Corrigendum 13 (1980) 3095.
[65] Z.-J. Wu, E.-J. Zhao, H.-P. Xiang, X.-F. Hao, X.-J. Liu, J. Meng, Phys. Rev. B 76 (2007), 054115.
[66] F. Mouhat, F.-X. Coudert, Phys. Rev. B 90 (2014), 224104.
[67] S. Singh, I. Valencia-Jaime, O. Pavlic, A.H. Romero, Phys. Rev. B 97 (2018), 054108.
[68] L. Rapp, B. Haberl, C.J. Pickard, J.E. Bradby, E.G. Gamaly, J.S. Williams, A. V. Rode, Nat. Commun. 6 (2015) 7555.
[69] L.-D. Zhao, S.-H. Lo, Y. Zhang, H. Sun, G. Tan, C. Uher, C. Wolverton, V.P. Dravid, M.G. Kanatzidis, Nature (London) 508 (2014) 373–377.
[70] P.-C. Wei, S. Bhattacharya, J. He, S. Neeleshwar, R. Podila, Y.Y. Chen, A.M. Rao, Nature (London) 539 (2016) E1–E2.