Machine learning prediction on intermetallic compounds with implemented virtual-center-atom structural descriptor

Zhenwei Li and Junjie Wang
State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi’an, Shaanxi, People’s Republic of China

ABSTRACT
A virtual-center-atom (VCA) structural descriptor was proposed to construct machine learning potentials which can be applied in studying alloy materials. The descriptor performs by describing the spatial features of atomic environments centered on a defined virtual atom within crystalline structure. Computational algorithms were developed to determine the position of virtual center atom. To verify the generality of this descriptor, crystal structures from both molecular dynamics trajectory and online accessible databases are utilized to check their machine learning (ML) performances. The challenge of the ML lies in increasing the transferability of the derived weight parameters to cover a broad range of structure distribution. Forces are found difficult to train the ML model under simple optimization methods. We thus suggest deep learning techniques for improving the applicability of the VCA potential. The descriptor is simple compared to other reported schemes and has promising applications in accelerating material discovery.

1. Introduction
The development of modern computing technology and density function theory (DFT) [1,2] has provided an unprecedented advancing power in material research. However, the simulated systems are always limited to those of several hundred atoms or those at a time scale of several nano seconds. This makes the observation or explanation of many interesting phenomena at meso- or multiscale level infeasible by full-accuracy DFT [3–12]. Material discovery is mainly hindered because of the prohibitive requirement in computing resources. Efforts to break the dilemma are attracted to the application with machine learning techniques [3–6, 13–20]. Artificial intelligence (AI) or machine learning (ML) has been hot research topics since quite early age in the computer history [21,22], however their significance in combining with material research was only uncovered lately [3–5,7,8,16–20,23–27]. The various associated ML exploration in material discovery can be found in relevant researches such as in the discovery of high-entropy alloys [27–29], high-Tc superconductors [24,25], molecular [20] or novel crystal structures [7,8]. First step of all for performing ML calculations is often to build a descriptor to describe the intrinsic features of the target material. The validity of a structural descriptor will largely determine the machine learning performance during the calculations. There are various representing and describing schemes in machine learning work to address a given material problem [3–5,7,8,13,14,18–20,24–27]. They can be generally categorized into two groups, either using chemical or physical quantities that are easy to
be measured or using structurally derived descriptors. There are several merits or criteria from which one can judge to use a descriptor in ML application, for example:

(i) The descriptor is bound to incorporating intrinsic symmetries of the structures, such as the translational, rotational and atomic permutational symmetries when dealing with crystal structures. Incorporation of symmetries in the descriptor will avoid the repeated data training with symmetrically identical structures and thus favor the ML efficiency immediately.

(ii) Uniqueness is another requirement for a descriptor to be applied in ML. Only with unique description for data structure can a descriptor be applied in material classifier or property predictor. Basically, a descriptor should be in line with the one-to-one or many-to-one correspondence between structure and calculated property.

(iii) Rigorousness has to be satisfied so that the structural complexity can be correlated with the dimensionality of the description. The structural descriptor is thus able to distinguish similarities or dissimilarities among structures in the database.

(iv) A much reduced deriving procedure will reduce costs in structure representation calculations and thus favor the overall ML efficiency.

(v) Direct correlation between the description elements and physical properties is advantageous for ML to perform. This not only reduces dimensionality of the representing features and will also expand the transferability of the prediction.

(vi) Adding physically or chemically measured properties into the descriptor will enhance the ML calculation [5,14], however, at a cost in increasing the overlapping in feature description, and thus an abstraction of principal features is often required.

(vii) An ML descriptor can be linked to the topology of materials and topology analysis methodology [30–32] can provide ideas for building descriptor, especially when both of them are targeting for novel structure design. However, the locality and simplicity requirement of descriptor have often to be justified in use.

(viii) With structurally derived descriptors, one can describe the ML potential energy surface as a function to the represented structure data, thus render the generation of new applicable potentials [3,4,13].

After satisfying the above criteria, along with the database calculated from first-principles, the structural varieties both from structural transformation or atomic permutations can be represented by descriptors and ML tasks can be established accordingly. One particular application of descriptor in machine learning calculation based on DFT data is to generate associated ML potentials, which by data-driven training, can speed up MD calculation to cover large-scale simulation. For those constructed ML potentials, the application of descriptors is especially significant because the regression from training and prediction based on potential energy surface (PES) requires higher resolution in distinguishing the features of structures in database. The machine learning potentials that have been so far applied in molecular dynamics simulations include Gaussian approximation potential (GAP) [3], neural-networks potential (NNP) [4] and spectral neighbor analysis potential (SNAP) [13] etc. Their latest development can also be found in [33–35]. In the GAP and SNAP potentials, bispectrum descriptor is adopted to describe the atomic environments which are directly associated with the machine learning target, i.e. local atomic energy \(E_i\), whose sum over all atoms will give the potential energy \(E_{\text{tot}}\). In the NNP, atomic-centered symmetry-functions are employed to construct descriptor for atomic configurations and the outcome further plays as input for the neural-network model of potential energy surface (PES) [4]. Specifically, machine learning training and prediction on DFT forces were also attempted in ref. [5] and are generally useful for ‘on-the-fly’ molecular dynamics simulation [10–12]. The descriptor for ML on force in Ref. [5] is in a simplified form while still preserving the symmetry invariance by describing atomic centered environment with feature matrix whose elements are inter-projections between intrinsic internal vectors. In a recent work by Kruglov et al [15], the same descriptor was adopted in building an energy-free machine learning force field for Aluminum. Thermodynamic and melting simulation with the derived force field reproduced the experimental and full ab initio results with very high accuracy. Though so far there are a number of choices for descriptor, there is still no such a descriptor that satisfies all the criteria under all application contexts in materials science. Using many different descriptors together can be promising, but to have an accurate prediction on structures, a lot of efforts have to be put into optimization from the tremendous number of candidate describing features before each machine learning run. The design of new descriptor is still a prerequisite and necessary step in ML approaches for solving material problems.

In this work, we focus on construction of machine-learning PES for intermetallic compounds of Nb3Si5 and Ti-Al. To this end, we defined a virtual center
atom (VCA) descriptor, with which we are able to describe energy-associated structural features by using the same technique as we applied for describing atomic environments in ML force calculation \[5,15\]. As a new descriptor among others, it is in simple form and very promising to be useful in material discovery. Computational algorithms are developed and implemented in our machine-learning package: ComputingVCApot. We will present further discussions based on our revealed computational results. The application of the descriptor renders a new type of machine-learning potential, i.e. VCA potential. Its generic applicability is tested upon database generated from first-principles molecular dynamics (FPMD) simulations \[36\] and databases that are from different resources.

2. Computational detail

2.1. Derivation of VCA descriptor

To start ML calculation, MD structural database was first transformed to VCA data via VCA procedures encoded in our ComputingVCApot code. For each structure, it is generically calculated by a two-step procedure:

(i) to determine the position of VCA.
(ii) to calculate its representing feature matrix based on VCA center.

By definition, VCA is a virtual center atom we can employ to build atomic-centered descriptor. It has no physical interaction with surrounding environments but its coordination has to be computationally derived. The latter part of the procedure is almost the same as that for the problem we met when performing ML on DFT forces solely \[5\]. The VCA position is playing a role of operating center for building the internal feature vectors (IFVs) descriptor.

2.2. Feature matrix calculation

In a structure, the surrounding atomic environments centered on an atom are largely determining most of the DFT properties and a descriptor is to get all the important features represented. Concerning such descriptors, there are already popular schemes like using bispectrum \[3\], symmetric functions \[4\] etc. They all take into account the crystal symmetries and locality features of the atomic neighboring environment but are quite from different aspects. For our study, we concentrated on using the IFVs to build descriptor for the geometry of a crystal structure \[5\]. Therefore, one can derive a feature matrix whose elements are composed of projected components between each pair of feature vectors. The IFVs are defined in Equation (1), where \((r_0, m)\) specifies a pair of possible hyper-parameters in the weight factor \(e^{-\left(\frac{r}{\delta}\right)^m}\), which we assumed on the \(i\)-th neighboring atom. Different pairs of \((r_0, m)\) will generate vectors overall giving a description of the structural features at different distance with respect to the center atom.

\[
v(r_0, m) = \sum_{i=1}^{\text{Nat}om} \vec{r}_i e^{-\left(\frac{r}{\delta}\right)^m}
\]

For compounds, we have to incorporate the distinct atom types [see supplemental material in reference \[5\]]. In Equation (2), a \(\delta(s_i - s)\) function is introduced to the neighbor position vectors \(\vec{r}_i, s\) indicates the atom type, such as \(s = 41\) or \(14\) for Nb and Si in compound Nb\(_3\)Si\(_3\), respectively (Table 1). As revealed, IFVs for binary compounds are doubled because the structural features now have to incorporate features for two atom types, and each of them is described using Equation (2). Similarly, a feature matrix is also calculated as inter-projections between all the IFVs, which is expressed in Equation (3). The \(V\) contains all the derived IFVs in matrix and \(A\) corresponds to matrix of their directional vectors. The feature matrix \(M\) is invariant to any rotational symmetric operation on the atomic configurations. This therefore generates a descriptor that is invariant to externally defined coordinates and can generally enhance the efficiency during ML computation on the compounds.

\[
v(r_0, m, s) = \sum_{i=1}^{\text{Nat}om} \vec{r}_i e^{-\left(\frac{r}{\delta}\right)^m} \delta(s_i - s)
\]

\[
M = V^T \bullet A
\]

Table 1. The \((r_0, m, s)\) parameters for generating IFVs descriptor for Nb\(_3\)Si\(_3\), \(N_{gpd} = 10\) for Nb \((s = 41)\) and \(N_{gpd} = 10\) for Si \((s = 14)\) were chosen as in the table. The data is included in grid.dat file as input for ComputingVCApot.

| \(r_0\) | \(m\) | \(s\) |
|---|---|---|
| 1 | 0.5 | 0.825 | 41 |
| 2 | 2.016 | 0.25 | 41 |
| 3 | 0.716 | 0.825 | 41 |
| 4 | 2.833 | 1.783 | 41 |
| 5 | 2.833 | 5.233 | 41 |
| 6 | 5.7 | 5.233 | 41 |
| 7 | 6.35 | 0.825 | 41 |
| 8 | 6.56 | 0.441 | 41 |
| 9 | 6.35 | 2.358 | 41 |
| 10 | 7.0 | 4.85 | 41 |
| 11 | 0.5 | 0.825 | 14 |
| 12 | 0.7 | 0.825 | 14 |
| 13 | 2.883 | 1.783 | 14 |
| 14 | 3.1 | 2.741 | 14 |
| 15 | 3.31 | 5.808 | 14 |
| 16 | 5.7 | 5.233 | 14 |
| 17 | 6.35 | 0.825 | 14 |
| 18 | 6.35 | 2.358 | 14 |
| 19 | 6.56 | 0.441 | 14 |
| 20 | 7.0 | 4.85 | 14 |
2.3. Exploration of descriptor application

The scheme of feature matrix descriptor in Equation (3) has been applied to machine learning on DFT forces in MD simulation and accelerative calculation can be performed in combination with on-the-fly QM database updating [5]. An energy-free machine-learning force field was also reported in literature based on the structural descriptor [15]. It proves to be an efficient descriptor for atomic-force-related machine-learning tasks. However, we hereby further extend the utility of the descriptor by applying the same descriptor to deal with machine learning of energy quantities. With an assumed virtual-center-atom, we found the IFV descriptor scheme can be conveniently borrowed in our ML tasks of PES energies and work along this line yields a simple ML potential, which we named VCA potential. In this work, after exploring significance of the VCA structural descriptor, we furthermore explore the related application of this new kind of ML potential.

2.4. Choice of ML model

Fully-connected neural networks (FCNN) are established by different layers of neurons. Each neuron can be treated by activation function and output to the next layer. The explicit activation functions can take function forms such as tanh, sigmoid, ReLU etc. Different layers of neurons are connected by adjustable weight parameters \( \{w^{(\text{layer})}_{ij}\} \), where, \( i \) and \( j \) denote the position in two adjacent layers and \( \text{layer} \) denotes the layer where the current computing neurons are in the neural-networks (NN) model. The sum of all outputs from the forth layer is adjusted with weight parameters and pass as inputs to next layer of neurons. The first input layer and the last output layer are manipulable while the other layers are treated as hidden layers. The machine-learning training is thus a process to obtain the optimized weight parameters by adopting optimization algorithms. ReLU has the simplest form with a stair-like slope function and is computationally robust in practice [23,24]. The other activation functions have relatively complicated forms and requires remarkable amount of computing time to obtain the explicit derivatives. We adopted a single-layer NN for our ComputingVCApot calculation. It is closest to linear-regression (LR) machine learning which favors calculation efficiency for different test reasons.

2.5. Data training algorithms

For single layer neural-networks, training with either back-propagation (BP) [21,37] or other optimization algorithms like steepest descent (SD) is robust enough. We incorporate two optimization algorithms in our code and adopt them during ML calculations for Nb-Si and Ti-Al. For multi-layer neural-networks model, the number of computed derivatives with respect to weight parameters increase rapidly by a factor proportional to the number of neurons, which causes the efficiency of data training to break down to its limit very quickly. The learning in ComputingVCApot was accomplished by first training the LR model with energy data and subsequently performing a finer optimization on the weight parameters to find the global minimum for both energy error and force error. The minimizing of ML error in Equation (4) have to be performed on both contribution from energy and force, while factor \( \lambda \) will adjust the ratio between them.

\[
f = f_{\text{energy}} + \lambda f_{\text{force}}
\]

Parallel calculation schemes are established by using a shifting \( \lambda \) parameter during each calculation run. The resulting weight parameters from each iteration will then be ordered on the same scale according to absolute values of their ML errors from either energy or force and will also be renewed whenever the two ordered lists are different. Therefore, the calculation always moves toward minimizing the difference between two lists of ordered weight parameters. The entire deep optimization calculation can stop till we get the weight parameters that meet the accuracy threshold. In ComputingVCApot, however, we note that the energy error term \( f_{\text{energy}} \) can be further categorized into two sources: the ML errors on pristine DFT energy data \( f_{\text{energy}, \text{DFT}} \) and ML errors on the augmented energy data \( f_{\text{energy}, \text{aug}} \). The latter are derived within harmonic approximation based on the information of calculated DFT forces but they are converted to energy data by a distortion on original structures along the force vectors. All the augmented structures and their corresponding energies are simply added to the entire training database.

2.6. Database generation

The DFT database of Nb3Si3 was constructed from structures and corresponding energy and force data calculated from first-principles MD simulation trajectories using VASP package [38]. NVT temperature was set to be 500 K, 1500 K and 2500 K to generate the database, and plane-wave energy cutoff was 520 eV while k-points were set to be \( 4 \times 4 \times 1 \). Ti-Al database was downloaded from the material platform in our group, which contains more structural varieties from MD trajectories of different temperatures, lattice deformations and different crystal types.

3. Results

3.1. VCA position by using geometric center

The algorithm to unveil the VCA center is nontrivial. As mentioned, the VCA atom has no physical
meaning in itself. In other words, there is no physical interaction between the assumed VCA atom and environments and thus no PES energy contribution accordingly. However, from the point of view of structural descriptor for crystalline compound, a VCA center will greatly simplify our computation during ML of energy-like quantities. One can preliminarily use geometric center as VCA center but will have increasing challenges to have a complete description for complex cases. For example, for multi-specie compounds, the geometric center will largely deviate from each other due to the distinct of atom types. Thus, an improbable VCA center may be solution to the question. We start from considering translational symmetries in the crystal structure. The primitive cell is the best unit to look for the VCA center which will be invariant to symmetry operations. However, external coordination of user-defined unit cell will depend on which atom we put on the origin. In other words, the user-defined unit cell can be shifted in 3-dimensional space. Our proposed procedure is illustrated in Figure 1, a number of user-defined unit cells are built by using each of the non-equivalent atomic sites as coordination origin respectively. The enclosed atoms within the primitive cell will have a geometric center but the valid VCA center for the entire crystal is generated by averaging on all these candidate centers. The uniqueness of the VCA center is thus achieved by averaging out the deviations of VCA centers caused from different choices of the origins, and using such an averaged VCA center will satisfy the criterion of descriptor. In the following work, we can perform our ML calculations on PES energies combining VCA center by this technique with the already developed IFVs descriptor.

3.2. Computationally determining VCA position

The prerequisite requirement for the VCA center is being unique within a cutoff radius $r_{\text{cut}}$. It will render much more uncertainties for the machine-learning database if the virtual center is not unique. Strictly speaking, the uniqueness can be ensured by setting a generating function $\mathcal{g}$ throughout the VCA computation. Under this condition, there are various ways to transform a number of atomic positions to a unique position. In the simplest case of elemental crystals, the $\mathcal{g}$ can take the form of summing up over all position vectors within a unit cell or atomic cluster Equation (5). Accordingly, the virtual center is the geometric center of the configuration, as mentioned above.

$$\mathcal{g} = \sum_{i=1}^{\text{Natim}} \bar{r}_i$$

For periodic system, the VCA descriptor calculation for structures is equivalent to finding the high-symmetric center of the crystal. The virtual center can be found via a four-step atomic labeling and calculation:

(i) To calculate the Wyckoff positions of the primitive cell under a Bravais lattice symmetry.
(ii) To build sublattices with origin on each of the atoms in Wyckoff position lists.
(iii) VCA centers are calculated as centers of the atomic configurations enclosed within each of the re-defined unit cells.
(iv) Practical VCA center is at the averaged position of the above VCA centers.

In the above procedures, atom $A_i$ belongs to the user-defined unit cell $P$, i.e. $A_i \subset P$. For each $P$, the enclosed atoms $\{A_i\}$ will generate a virtual center while for $N_p$ sublattices, there are correspondingly $N_p \times N_{\text{energy}}$ pieces of energy data. Note that, these $N_p$ configurations have the same energy, and thus they are energetically degenerated. The energetically degenerated database is actually enlarged to $N_p \times N_{\text{energy}}$ in size, which means that two close configurations having close centers are highly energetically equivalent. In this case, the IFVs have to be sensitive enough to distinguish the difference of two very close structures at such a resolution. The subtleness of this case will accordingly require a more sophisticated ML model. We normally adopt the averaged center to accomplish the task of ML on energies. This means we have $N_{\text{energy}}$ pieces of represented structures and energies, as shown in Figure 2. By using the averaged VCA center, we will put more emphasize on the training and learning efficiency.

Using the geometric center is the most straightforward approach to calculate the VCA position. The averaged VCA position based on a number of such defined VCA centers are less robust but still preserving a good efficiency, as symmetries of the crystal can often bring simplification to the calculations. These two schemes of VCA are explicitly implemented in
Computing VCA pot and denoted as simple VCA and shift-cell VCA, respectively. In principle, combination of two unique VCA centers will yield a new definition of VCA, thus there are other possible candidate schemes. However, systematic validation with the basic ones will often provide an estimation for the upper limit of ML accuracy that the VCA descriptors can achieve. We thus only focus on the comparison calculation between shift-cell and simple VCA schemes in the following work.

3.3. Validation of VCA schemes

We start from testing the VCA ML scheme on α-Nb₅Si₃ structures. It is a body-centered tetragonal phase of Nb₅Si₃ (space group I4/mcm, and lattice constants  a = 6.56Å, c = 11.88 Å) and most stable at room temperature. Due to the importance of Nb-Si-based alloys in aerospace industry, the study of these kinds of materials has never been trivial [39]. We choose it as our ML target material to test the developed algorithms. ML database of Nb₅Si₃ was generated from first-principles MD by setting T from room temperature to 2500 K, calculated within a unit cell incorporating 16 atoms. No melting was found during the process of database generation. The setting of such high temperature is to ensure the structural space is reasonably sampled by the trajectory structures of the MD. As mentioned above, for α-Nb₅Si₃ database generated from MD, we can derive the VCA center via either shift-cell VCA method or simple VCA method, both of which will simultaneously move with the other atomic sites along the MD trajectory. In case of database generated from general sources or by users for different purposes, where the unit cells have more arbitrary choices, the shift-cell method is more reliable because it can smooth out the influence from such an external choice on a unit cell. We compared the ML results from these two approaches and the results are given in Figure 3. For the case of Nb₅Si₃, after LR training on the database, the ML accuracy is found satisfactorily good for a broad range of structural variation. From the error curves, we however find that the simple scheme has better training accuracy both for the force-vector augmented database (augmented case) and the pristine DFT energy database (non-augmented case). The shift-cell scheme is less accurate than simple scheme in the training even though the optimal accuracy of both are quite close. The reason may be rooted in that there is some substantial structural complexity introduced by shifting cell operations for MD structures. The absolute training errors are relatively large for TiAl, because more varied and transformed unit cells are calculated and incorporated in the database. The shift-cell can be advantageous in this case. There is notable difference between being augmented and non-augmented for the database. Also in Figure 3, the entire training means training the ML model with all sources of data while partial training means that a prediction calculation was further performed retrospectively on the pristine DFT energy database after the entire training. In contrast, for TiAl database, both using the augmented database and entire training method now favors the accuracy significantly and shift-cell VCA training has better achieved accuracy than simple VCA.

3.4. Influence of r_cut factor

After determining the VCA center either by the shift-cell method or simple VCA method, a proper choice for the cutoff radius r_cut of the atomic environment is important. It is upper limit that the r_0 value can take and therefore is directly connected to construction of the IFVs and also the descriptor feature matrix.
For machine learning of DFT forces, the cutoff radius $r_{\text{cut}}$ can be firstly estimated by DFT calculations on isolated atomic configurations formed with surrounding atoms centered on a given atom. Until a large enough $r_{\text{cut}}$ is adopted, the DFT force on the center atom will not converge and differs from that it is in the periodic crystal. The minimal $r_{\text{cut}}$ for this convergence to happen is normally considered to be the best $r_{\text{cut}}$ for our IFVs descriptor. After a $r_{\text{cut}}$ is chosen, IFVs calculations will only be performed on the neighboring atoms within the cutoff distance. In contrast, during our VCA calculation, the converging $r_{\text{cut}}$ has to be more enlarged compared to the case of ML on DFT forces. This is to ensure that we can describe features not only surrounding the VCA center but also incorporating feature description centered on all relevant atoms within the unit cell. Generally speaking, the features surrounding the center atom with force should be less complex than that for building energy descriptor to describe crystals. For the same material, we therefore initially adopt a larger $r_{\text{cut}}$ for VCA descriptor while the explicit value for $r_{\text{cut}}$ can be determined numerically using the convergence test procedures. To explicitly see the contributions from choosing different $r_{\text{cut}}$ in VCA calculation, we plotted the ML training results on $\alpha$-Nb$_2$Si$_3$ database in Figure 4. the cases of $r_{\text{cut}} = 6$ Å and 15 Å are calculated in comparison. In the Figure, for each given $N_{\text{grid}}$, there are 40 sampled griddat (abbreviation for grid.dat file) to perform independent linear-regression ML runs. The training errors will be calculated by comparison of ML results with the DFT benchmark. It can be generally seen that larger $r_{\text{cut}} = 15$ Å is generating more accurate training model than $r_{\text{cut}} = 6$ Å. Because larger $r_{\text{cut}}$ was chosen, we have to correspondingly increase the grid mesh we use to calculate the IFVs as well. In both cases, the overall training errors (average and minimal values) are monotonically down decreasing along with increasing $N_{\text{grid}}$.

The detailed error distribution by training with $r_{\text{cut}} = 15$ Å is also plotted in Figure 5. Apart from decreasing tendency, the errors have a broadened distribution as we increase $N_{\text{grid}}$ for generating the IFVs and ultimately the feature matrix. The solely increasing of $N_{\text{grid}}$ can not ensure that the ML model will be more generic, as too large $N_{\text{grid}}$ will increase the dimension of feature matrix simultaneously which may result into overfitting rapidly. Therefore, we have to cease from using too large $N_{\text{grid}}$ and instead use moderate $N_{\text{grid}}$ to balance the requirement between training and prediction accuracy. The numerical solution in ComputingVCApot for solving overfitting problem is by taking a further step to perform a finer optimization on hidden weight parameters or combine with some deep-learning techniques (see Equation (4)).

### 3.5. Evaluation of VCA potential

The LR machine learning calculation will minimize the loss on energy fitting as much as possible. However, the minimization process will not always favor the ML on atomic forces, which means our model derived above can only reproduce the energies and has inferior prediction on forces. To be applied in MD simulation, the ML forces of our VCA potential (VCApot) have to satisfy requirement on accuracy as well. The finer optimization in Equation (4) uses the errors on force as regulated term, and global minimum for both energies and forces are sought for in a parallel scheme. Automatic ranking and ordering of training errors are
Figure 4. The mean (MEAN), variance (VAR), maximum (MAX) and minimum (MIN) of training errors are calculated for each given \( N_{\text{grid}} \). Symbols in blue are for smaller \( r_{\text{cut}} = 6 \text{ Å} \) and those in red are for larger \( r_{\text{cut}} = 15 \text{ Å} \) while solid and dotted lines are marking the average and minimum of the error distributions along the axis of \( N_{\text{grid}} \). Four subplots give the results on mean, variance, maximum and minimum calculation, respectively. All training errors are in unit of eV/atom but are plotted with their natural logarithm in the Figure.

Figure 5. The tendency for increasing the number of representing internal feature vectors in the VCA calculation on Nb$_5$Si$_3$ database. Mean (MEAN), variance (VAR), maximum (MAX) and minimum (MIN) of ML errors are depicted in each of the subplots, respectively. Variance at each \( N_{\text{grid}} \) is caused by using different combination of \( (r_0, m, s) \) parameters. All errors are in unit of eV/atom but are plotted with their natural logarithm. In the figure, the red lines are averaged value along each given \( N_{\text{grid}} \) and \( N_{\text{grid}} \) also correspond to the number of IFVs generated.
performed for regenerating possible hidden weight parameters at each run for performing iterative finer optimization. Combining with our described deep learning strategy, we established the approach to generate reliable machine learning VCApot, which is incorporated in our computing Python package: Computing VCApot.

The construction of PESs for multi-specie intermetallic compounds is challenging because there is large amount of structural variation in the presence of different element types. Under VCA, we however can see that, the variation is largely reduced by taking into account of the intrinsic symmetries of structures during extracting the ML input features. Selective neighboring atomic configurations are constructed for describing each of atom types in compounds via using δ function and the overall features are derived from intrinsic projections between the IFVs. For α-Nb₅Si₃ database, one of such generated VCApot has tested accuracy results depicted in Figure 6. In case of accurate ML energies, the derivation of correct force vectors is still a challenging task. Currently, we can only use the developed technique to generate VCApot for 100–200 MD structures, and relevant algorithms are still undergoing improvement to describe more general structure types. Even though, we can anticipate the promising application of applying the VCA descriptor scheme.

4. Summary and discussion

In conclusion, we developed a VCA structural descriptor which can undertake the prerequisite feature engineering task for machine-learning on metallic compound structures. Different from descriptors for machine learning of forces on local atomic environment, this descriptor involves the determination of a virtual-center-atom position based on computational procedures. We proposed two approaches including simple VCA and shift-cell VCA schemes and investigated their machine-learning performance on Nb₅Si₃ and Ti-Al database. Within the ML framework, we are able to generate a VCA potential that is promising to be used in MD simulations. To satisfy required accuracy threshold, deep-learning or finer optimization techniques were attempted accordingly. As a new proposed structural descriptor, we think that the proposed VCA schemes can make use of both the advantages from ML and structural descriptor in addressing material problems. The proposed descriptor may thus find general applications in ML study of materials while providing useful hints for future exploration of other similar ML tasks on DFT databases.
Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was financially supported by the National Nature Science Foundation of China (Grant No. 51872242 and 52111530033); National Natural Science Foundation of China [52111530033]; the Fundamental Research Funds for the Central Universities [D5000220172]

References

[1] Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev B. 1964;136(3B):B864.
[2] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation elects. Phys Rev. 1965;140(AA):A1133–A1138.
[3] Bartók AP, Payne MC, Kondor R, et al. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett. 2012;104(13):136403. DOI:10.1103/PhysRevLett.104.136403
[4] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett. 2007;98(14):146401.
[5] Li Z, Kermode JR, De Vita A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. Phys Rev Lett. 2015;114(9):096405.
[6] Kermode JR, Peralta R, Li Z, et al. Proc Mater Sci. 2014;3:1681. DOI:10.1016/j.mspro.2014.06.271.
[7] Nyshadham C, Rupp M, Bekker B, et al. Machine-learned multi-system surrogate models for materials prediction. Npj Comput Mater. 2019;5(1):51. DOI:10.1038/s41524-019-0189-9
[8] Ziletti A, Kumar D, Scheffler M, et al. Insightful classification of crystal structures using deep learning. Nat Commun. 2018;9(1):2775. DOI:10.1038/s41467-018-05169-6
[9] Caccin M, Li Z, Kermode JR, et al. A framework for machine learning augmented multiscalar atomistic simulations on parallel supercomputers. Int J Quantum Chem. 2015;115(16):1129–1139. DOI:10.1002/qua.24952
[10] Bernstein N, Kermode JR, Csányi G. Hybrid atomistic simulation methods for materials systems. Rep Prog Phys. 2009;72(2):026501.
[11] Csányi G, Albaret T, Payne M, et al. “Learn on the fly”: a hybrid classical and quantum-mechanical molecular dynamics simulation. Phys Rev Lett. 2004;93(17):175503. DOI:10.1103/PhysRevLett.93.175503
[12] Kermode JR, Albaret T, Sherman D, et al. Low-Speed fracture instabilities in a brittle crystal. Nature. 2008;455:1224.
[13] Wood MA, Thompson AP. Extending the accuracy of the SNAP interatomic potential form. J Chem Phys. 2018;148(24):241721.
[14] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B. 2013;87(18):184115.
[15] Kruglov I, Sergeev O, Yanilkin A, et al. Energy-Free machine learning force field for aluminum. Sci Rep. 2017;7(1):8512. DOI:10.1038/s41598-017-08455-3
[16] Snyder J, Rupp M, Hansen K, et al. Finding density functionals with machine learning. Phys Rev Lett. 2012;108(25):253002. DOI:10.1103/PhysRevLett.108.253002
[17] Szlacht WJ, Bartók AP, Csányi G. Accuracy and transferability of gaussian approximation potential models for tungsten. Phys Rev B. 2014;90(10):104108.
[18] Sh apeev AV. Moment tensor potentials: a class of systematically improvable interatomic potentials. Multiscale Model Simul. 2016;14(3):1153–1173.
[19] Botu V, Ramprasad R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. Int J Quantum Chem. 2015;115(16):1074–1083.
[20] Rupp M, Tkatchenko A, Müller KR, et al. Rupp et al. Reply:. Phys Rev Lett. 2012;109(5):059802. DOI:10.1103/PhysRevLett.109.059802
[21] Haykin S. Neural networks and learning machines. China: China Machine Press; 2009.
[22] Mackay DJC. Information theory inference and learning algorithms. UK: Cambridge University Press; 2003.
[23] Zeng S, Li G, Zhao Y, et al. Machine learning-aided design of materials with target elastic properties. J Phys Chem C. 2019;123(8):5042–5047. DOI:10.1021/acs.jpcc.9b01045
[24] Zeng S, Zhao Y, Li G, et al. Atom table convolutional neural networks for an accurate prediction of compounds properties. Npj Comput Mater. 2019;5(1):84. DOI:10.1038/s41524-019-0223-y
[25] Stanev V, Oses C, Kusne AG, et al. Machine learning modeling of superconducting critical temperature. Npj Comput Mater. 2018;4(1):29. DOI:10.1038/s41524-018-0085-8
[26] Li H, Wang Z, Zou N, et al. Deep neural network representation of density functional theory Hamiltonian, 2021; DOI: 10.48550/arXiv.2104.03786.
[27] Hart G, Mueller T, Toher C, et al. Machine learning for alloys. Nat Rev Mater. 2021;6(8):730–755. DOI:10.1038/s41578-021-00340-w
[28] Islam N, Huang W, Zhuang HL. Machine learning for phase selection in multi-principal element alloys. Comput Mater Sci. 2018;150:230–235.
[29] Huang W, Martin P, Zhuang HL. Machine-learning phase prediction of high-entropy alloys. Acta Mater. 2019;169:225–236.
[30] Akhmetshina TG, Blatov VA, Proserpio DM, et al. Topology of intermetallic structures: from statistics to rational design. Acc Chem Res. 2018;51(1):21–30. DOI:10.1021/acs.accounts.7b00466
[31] Nicholas TC, Alexandrov EV, Blatov VA, et al. Visualization and quantification of geometric diversity in metal–organic frameworks. Chem Mater. 2021;33(21):8289–8300. DOI:10.1021/acs.chemmate r.1c02439
[32] Mitrofanov A, Andreadis N, Korolev V, et al. A search for a DFT functional for actinide compounds. J Chem Phys. 2021;155(16):161103. DOI:10.1063/5.0067631
[33] Behler J. Perspective: machine learning potentials for atomistic simulations. J Chem Phys. 2016;145(17):170901.

[34] Szlachta WJ, Bartók AP, Csányi G. Accuracy and transferability of Gaussian approximation potential models for tungsten. Phys Rev B. 2014;90(10):104108.

[35] Wood MA, Cusentino MA, Wirth BD, et al. Data-Driven material models for atomistic simulation. Phys Rev B. 2019;99(18):184305. DOI:10.1103/PhysRevB.99.184305

[36] Car R, Parrinello M. Unified approach for molecular dynamics and density-functional theory. Phys Rev Lett. 1985;55(22):2471.

[37] Linainmaa S The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors (Masters thesis). Univ. Helsinki; 1970.

[38] Kresse G, Furthmuller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B. 1996;54(16):11169.

[39] Bewlay B, Jackson M, Subramanian P, et al. A review of very-high-temperature Nb-Silicide-based composites. Metall Mater Trans a. 2003;34A(10):2043. DOI:10.1007/s11661-003-0269-8