공학박사 학위논문

Crystal structure prediction using machine learning potentials and its applications

기계학습 퍼텐셜을 이용한 결정구조예측과 그 응용

2022년 2월

서울대학교 대학원
재료공학부
강성우

Crystal structure prediction using machine learning potentials and its applications

기계학습 퍼텐셜을 이용한 결정구조예측과 그 응용

지도교수 한 승 우

이 논문을 공학박사 학위논문으로 제출함

2022년 2월

서울대학교 대학원

재료공학부

강 성 우

강성우의 박사 학위논문을 인준함

2021년 12월

위 원 장	강 기 석	(인)
부위원장	한 승 우	(인)
위 원	이 원 보	(인)
위 원	한 상 수	(인)
위 원	임 진 오	(인)

Abstract

Crystal structure prediction aims to find the ground-state structure in a given composition. This is of great interest as it can establish a list of all synthesizable materials prior to experiments. However, the main challenge in predicting crystal structure comes from the efficiency of the algorithm: the NP-hardness of the problem and the high cost of density functional theory, which is employed as a structure optimizer and an energy evaluator, limit the widespread use of the algorithm in searching complex multinary systems. To accelerate the speed of crystal structure prediction, there have been several attempts to employ machine learning potentials as a surrogate model of density functional theory calculations. However, constructing the training set is not straightforward because prior knowledge of the configurations is not available before making predictions. Previous researches employed random sampling and on-the-fly sampling methods to train machine learning potentials but did not achieve enough efficiency and accuracy to be utilized in multinary systems.

In this dissertation, we develop the crystal structure prediction program using neural network potentials as the surrogate model of density functional theory calculations. Our main idea is to construct the training set with the disordered structures sampled from molecular dynamics simulations. The energies calculated by trained potentials show a good correlation with the energies calculated by density functional theory calculations, which indicates that the neural network potential can be a hi-fidelity surrogate model for crystal structure prediction. Then, we develop the crystal structure prediction method by optimizing algorithms for constructing training sets, training neural network potentials, and searching structures with evolutionary algorithms. The developed program is tested on the experimental

database and theoretical structures predicted by other crystal structure prediction methods. The tests show that the developed method can identify the global minimum in most cases at a reasonable computational cost. Using the developed method, we are now discovering the missing ternary metal oxides and Li superionic conducting oxide materials. By harnessing the accuracy and efficiency of neural network potentials, this dissertation will pave the way to the wide material discoveries in various research fields.

Keyword : crystal structure prediction, machine learning potential, neural network potential
Student Number : 2016-20761

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview of crystal structure prediction

In modern times, the discovery of new inorganic materials catalyzes the advancement of diverse technologies such as $InGaZnO_4$ for transparent flexible electronics [1], $Li_{10}GeP_2S_{12}$ for solid-state batteries [2], $Cs_2AgBiBr_6$ as photovoltaics [3], and $La[O_{1-x}F_x]FeAs$ [4], H_3S [5] and LiH_{10} [6,7] as high-T_c superconductors. Indeed, ~200,000 materials have been registered in the Inorganic Crystal Structure Database (ICSD), which contains information of most the synthesized inorganic materials, and its size increases by ~5,000 every year [8].

However, such experimental throughput for discovering new materials is not enough to complete the list of material database: based on a rough estimate, only ~16% of ternary materials and ~1% of quaternary materials are known among the possibly synthesizable materials list [9], and only ~1% of ternary materials and ~0.05% of quaternary materials are newly discovered every year.

Computational methods based on density functional theory (DFT) are widely employed in material searches because this can predict materials properties much faster than experiments without any prior knowledge [10]. In DFT calculations, crystal information of materials is the essential input that defines the Hamiltonian of

the given system. However, crystal structures of as-yet-synthesized materials are not available, so most of the experimental researches focus on screening already existing materials in the current database(e.g., ref. [11,12]).

To resolve this issue, one needs a crystal structure prediction (CSP) algorithm, which is a computational method to predict the crystal structures of solids from first principles [13]. In principle, an efficient CSP algorithm can identify all the synthesizable but as-yet-synthesized materials. Therefore, developing a practical CSP algorithm has been a dream of theoretical materials scientists. However, CSP is an NP-hard problem with no general solution [14].

Two kinds of CSP algorithms have been developed to mitigate this frustration: the first one is a data-driven method and the other one is a heuristic approach. The first systematic data-mining method is developed by Fischer et al. [15]. In this approach, the correlation between structures in terms of constituent elements is learned from the experimental database by machine learning. A similar algorithm has been developed by Hautier et al., which uses the machine learning model that learns the substitutional probability for element pairs for the given composition [16]. Within these or other data-driven methods, the new metal nitrides [17], 18-ABX compounds [18], and Li-ion batteries [19] are discovered followed by the experimental verification. However, this approach is not possible to find a material with a new prototype that is not in the database. In addition, the biased elemental frequencies in the experimental database might risk deficient prototypes in certain materials groups. In specific, oxygen is the most abundant element in the ICSD of which frequency is 3 times higher than the next most abundant elements (Fe, Si, and S). Therefore, the prototype database for other compounds than oxides may be deficient to construct a reliable statistical model.

To directly explore the configurational space instead of using statistical relations, several heuristic approaches, such as random structure sampling [20,21], metadynamics [22], simulated annealing [23], minima hopping [24], basin

hopping [25], particle-swarm optimization (PSO) [26,27], and genetic algorithm [28], are developed. The most popular approaches are the genetic algorithm developed by Oganov et al. (USPEX code) [29], PSO method developed by Ma et al. (CALYPSO code) [30] and random sampling developed by Pickard et al. (AIRSS code) [20]. These algorithms often accompany thousands or more DFT calculations which takes large computation time. Therefore, diverse techniques, such as random structure generation using symmetry operators, and antiseed option are developed for the efficient search of the ground-state structures in the short generations limited by the speed of DFT calculations [31]. With this method, new materials are successfully discovered such as the high-pressure phase of transparent sodium [32], the unusual stoichiometry of sodium chlorides [33], and H-based novel high-T_c superconductors [34–36]. However, the current researches based on CSP are often limited to short evolutionary generations [18,37], because the computational cost of DFT-based CSP is still demanding.

Recently, machine learning potentials are gaining much attention because they can be a surrogate model of DFT with much lower computational cost and the linear scaling with the number of atoms in contrast to the cubic scaling of DFT calculations [38–40]. Machine learning potentials learn structure-energy relations from DFT calculations via flexible functional forms. Therefore, machine learning potentials require the construction of a training set which consists of DFT calculation results on reference structures that contains every configurational information of target simulation. However, in the case of CSP, this encounters a significant challenge, because of the prior information of target simulation is not known before one makes the prediction. To resolve this, previous studies utilized random sampling or on-the-fly sampling method for training machine learning potentials for CSP [41–43]. The developed potentials were tested on unary crystals and clusters. However, these methods are less effective in more complicated multinary materials because they unnecessarily gather high-energy structures that are irrelevant to the local

minimums around the ground-state. Consequently, there is an urgent need for efficient strategies for training reliable machine learning potentials during CSP procedures and new search algorithms that is optimized for machine learning potentials.

1.2 Goal of the dissertation

The main goal of this dissertation is to develop an efficient and accurate CSP method using machine learning potentials. The CSP algorithm consists of two parts, global optimization, and local relaxation. In inorganic materials, the local relaxation part takes most of the time because it is conducted by DFT calculations. We first develop an efficient method to train neural network potentials (NNPs) employed for local relaxation and energy evaluation of structures that appear during the global minimization. The key strategy is to use disordered structures, which are generated from DFT-based melt-quench-annealing simulations, as a training set of NNP. We demonstrate that the generated potentials can be a high-fidelity surrogate model of DFT in CSP of multinary systems.

When applying NNPs in CSP, we find that the conventional global optimization algorithms are inefficient for NNP-based CSP because the conventional algorithms are only optimized for DFT calculations. Therefore, we develop the CSP code that is optimized for NNP. In addition, we additionally optimize the DFT-MD procedure, training methods, and quality monitoring of NNP. Based on these algorithms, we develop the code named SPINNER (Structure Prediction of Inorganic crystals using Neural Network potentials with Evolutionary and Random searches). We conduct two tests to evaluate the performance of the program: the first one is a blind test on the experimental structures and the other one is a benchmark test on theoretical structures predicted by other CSP algorithms. In both tests, it is confirmed that SPINNER can mostly identify a global minimum for a given composition within a reasonable computational cost.

Finally, we present two practical applications of CSP: identifying missing ternary oxides and discovering novel Li superionic conductors. We establish the strategies to efficiently search large materials space using SPINNER.

1.3 Organization of the dissertation

The dissertation consists of five chapters. Chapter 1 provides an introduction, which presents an overview of crystal structure predictions, the goal of the dissertation, and the organization of this dissertation. Chapter 2 gives the theoretical background on density-functional theory, machine learning potentials, and crystal structure prediction. Chapter 3 provides the discussion on the development of the neural-network potentials and the optimization of the crystal structure prediction algorithm. With the developed code, two performance tests are conducted. In Chapter 4, we discuss the practical applications of CSP. Finally, in Chapter 6, we summarize and conclude the dissertation.

Chapter 2

Theoretical background

2.1 Density functional theory calculations

2.1.1 Born-Oppenheimer approximation

When trying to calculate the material properties within atomic level without any empirical knowledge, one requires to solve the Schrödinger equation, which is the basic governing equation describing all physics in atomic scale. For the problems in materials science, systems consist of multiple nuclei and electrons. In this case, the Schrödinger equation is written as:

$$H\Psi = E\Psi, \tag{2.1}$$

where the Hamiltonian operator H is represented as:

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{r}_I - \mathbf{r}_J|} + \sum_{i,J}\frac{Z_J e^2}{|\mathbf{r}_i - \mathbf{r}_J|}. \tag{2.2}$$

Here, *i, j* are the index for electrons, *I, J* are the atomic index, **r** is the position of

electrons and atoms, *Z* is the charges of electrons and atoms, and *M* is the nuclear mass. The first two terms in equation (2.2) is the kinetic energy of electrons and nuclei, and the next terms correspond to the electrostatic interactions of electron-electron, nucleus-nucleus, and electron-nucleus, respectively.

Solving the equations (2.1) and (2.2) is a many-body problem which has no general solution except for the case of a hydrogen atom. Density functional theory (DFT) applies several approximations to provide reasonable first-principles solutions. The first one is the Born-Oppenheimer approximation [44]. In this approximation, the nuclei is assumed to be fixed while only the movements of electrons are considered. This is because the electron mass is orders of magnitude smaller than the nucleus mass. The kinetic energy of the nuclei approaches to zero, and the nuclei are just considered as a source of charge. Mathematically, equation (2.2) changes into:

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_i V_{\text{ext}}(\mathbf{r}_i) + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{r}_I - \mathbf{r}_J|}, \tag{2.3}$$

where $V_{\text{ext}}(\mathbf{r}_i)$ is the external potential applied to the electron from the nucleus charges, which is defined as:

$$V_{\text{ext}}(\mathbf{r}_i) = \frac{1}{2}\sum_I \frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{r}_I|}. \tag{2.4}$$

The last term in equation (2.3) is the constant, so one has to consider the first three terms to solve this equation. Therefore, one can rewrite Hamiltonian of the system as follows:

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_i V_{\text{ext}}(\mathbf{r}_i) + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{2.5}$$

2.1.2 Hohenberg-Kohn theorem

The key principle to solve equation (2.5) is the Hohenberg-Kohn theorem [45]. Hohenberg-Kohn theorem consists as follows:

(1) For electrons under external potential, $V_{\text{ext}}(\mathbf{r})$, the ground state is uniquely given as a function of electron density ($\rho(\mathbf{r})$).
(2) A universal functional for energy, $E[\rho(\mathbf{r})]$, exists as a function of $\rho(\mathbf{r})$, where the ground state energy is the global minimum of this functional in density, ρ_0.

The Hamiltonian in equation (2.5) is defined by ρ_0. Therefore, with this theorem, the energy functional is written as follows:

$$E[\rho(\mathbf{r})] = \int V_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \quad + F[\rho], \tag{2.6}$$

where,

$$F[\rho] = T[\rho] + V_{\text{ee}}[\rho]. \tag{2.7}$$

Here, $T[\rho]$ is the kinetic energy and $V_{\text{ee}}[\rho]$ is the electron-electron interaction term.

According to the Hohenberg-Kohn theorem, the wave functions (i.e., all properties) are completely given by ρ_0. When assuming there exists no degeneracy, this can be proven as follows: suppose there are two different external potentials $V_{\text{ext}}(\mathbf{r})$ and $V'_{\text{ext}}(\mathbf{r})$ which lead to the same ρ_0. These two potentials give two different Hamiltonians, H and H', and corresponding different wave functions, ψ and ψ'. Then, energy of the system, *E* is given by:

$$E = \langle\psi|H|\psi\rangle, \tag{2.8}$$

and also the following relation is satisfied because ψ is the ground state of Hamiltonian H:

$$\begin{aligned} E = \langle\psi|H|\psi\rangle &< \langle\psi'|H|\psi'\rangle = \langle\psi'|H'|\psi'\rangle + \langle\psi'|H - H'|\psi'\rangle \\ &= E' + \langle\psi'|H - H'|\psi'\rangle = E' + \int [V_{\text{ext}}(\mathbf{r}) - V'_{\text{ext}}(\mathbf{r})]\rho_0(\mathbf{r})d\mathbf{r}. \end{aligned} \tag{2.9}$$

Conversely, the below relation satisfies:

$$E' < E + \int [V'_{\text{ext}}(\mathbf{r}) - V_{\text{ext}}(\mathbf{r})]\rho_0(\mathbf{r})d\mathbf{r}. \tag{2.10}$$

Equations (2.9) and (2.10) are contradictory to each other, so this means that the assumption that two different wave functions exist for the same ρ_0 is wrong.

Despite the fact that Hohenberg-Kohn theorem proves that $E[\rho(\mathbf{r})]$ exists, it does not demonstrate any actual form of the functional. Chapters 2.1.4 provides descriptions to make a reasonable approximations of $E[\rho(\mathbf{r})]$.

2.1.3 Kohn-Sham equation

To effectively address the difficulties in the many-body problem of solving Schrödinger equation, Kohn and Sham assumed that the electron density $\rho(\mathbf{r})$ of N electrons can be divided into the sum of orbital charges of each electron [46]:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2. \tag{2.11}$$

The one-electron orbital $\phi_i(\mathbf{r})$ is called Kohn-Sham orbital. This is the mean-field approximation which transforms many-body problem in to the set of one-body problems. Transforming equation (2.5) using equation (2.11), Kohn-Shan equation is given as:

$$\left(-\frac{\hbar^2}{2m_e}\nabla_i^2 + V_{\text{eff}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \tag{2.12}$$

where i is the atomic index, ε_i is the Kohn-Shan eigenvalue of ith electron, and $V_{\text{eff}}(\mathbf{r})$ is the effective potential written as:

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r}). \tag{2.13}$$

Here, $V_{\text{Hartree}}(\mathbf{r})$ is the Hartree energy which corresponds to the Coulomb interaction term of an electron with itself:

$$V_{\text{Hartree}}(\mathbf{r}) = \frac{\mathrm{e}^2}{2}\int d\mathbf{r}d\mathbf{r}'\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\ . \tag{2.14}$$

$V_{\text{xc}}(\mathbf{r})$ is the exchange-correlation energy which includes all the corrections from

the many-body interactions among electrons. The exact formulation of the functional is not known. Therefore, the approximated functionals are used, which will be discussed in the next subsection.

When trying to solve equation (2.12), one encounters the self-consistency problem between the electron density and the wave function. To resolve this problem, iterative method is used: First, the electron density is guessed from the scratch. Then, the wave function can be determined by solving equation (2.12). With wave function, the electron density is update. And then, get the wave function again by solving equation (2.12). This iterations are performed until the electron density is consistent to the wave function.

2.1.4 Exchange-correlation energy

In this Chapter, we discuss the exchange-correlation energy (E_{xc}). The exact form of E_{xc} is now known, so one needs to use an approximate form. The simplest approximation of E_{xc} is the local density approximation (LDA) [46], which assumes that the local exchange-correlation energy is same as that of the free electrons with the same electron density. Therefore, LDA form of E_{xc} is written as:

$$E_{\mathrm{XC}}^{\mathrm{LDA}} = \int d\mathbf{r}\rho(\mathbf{r})\epsilon_{\mathrm{XC}}[\rho(\mathbf{r})]. \tag{2.15}$$

Generally, LDA provides reasonable results for the systems with slowly varying charge density. However, for systems with an electronic structure that significantly deviate from the uniform electron gas, LDA shows a large error. To refine this, generalized gradient approximation (GGA) E_{xc} considers the gradient of the electron density in the functional form [47,48]:

$$E_{\mathrm{XC}}^{GGA} = \int d\mathbf{r}\rho(\mathbf{r})\epsilon_{\mathrm{XC}}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]. \tag{2.16}$$

GGA tends to provide more reliable results compared to LDA. However, both GGA and LDA functionals show a non-negligible error. For instance, both functionals severely underestimate the band gap by 30-40%. This error is originated from the fact that the functionals do not consider the electron self-interaction energy and also derivative discontinuity at the integer number of electrons. Beyond-DFT methods such as GW approximation [49] or hybrid functionals [50] are known to partly address this problems. However these functionals take orders of magnitude longer time compared to LDA or GGA. Recently, SCAN functional is developed that can provide accurate energy orderings and structural properties with much lower

computational cost compared to GW other advanced methods [51,52]. However, SCAN functional still underestimates the bandgap [53,54].

2.2 Neural network potential

While DFT calculations provides accurate prediction of various properties, this demands a substantial computational cost. Recently, machine learning potentials are gaining much attention, as they can achieve similar accuracy to the DFT calculations but at a much higher speed. In addition, linear scaling of machine learning potentials with the number of atoms makes them more advantageous in modeling large systems. The first practical machine learning potential developed is Behler-Parinello type neural network potential (NNP) [38]. Since then, several types of machine learning potentials, such as Gaussian approximation potential (GAP) [40] and moment tensor potential (MTP) [55], have been developed. These potentials are successfully applied to diverse studies such as modeling amorphous structures [56] and describing the phase transition of hydrogen liquids [57].

Among various machine learning potentials, NNP and GAP are most studied and applied to various kinds of simulations. NNP is based on an artificial neural network that can describe any functional form when the number of parameters is sufficient in principle [58]. Therefore, they are suitable for learning diverse materials systems. Neural network potentials require a large amount of reference training set, and the training requires a high computational cost. On the other hand, GAP is based on the kernel-based model that consists of the linear combination of kernel functions that represent each reference point in the training set. Mathematically, this corresponds to the artificial neural network with 1 layer, but with a much larger number of nodes. The training of kernel-based models is directly performed by matrix algebra, so the training cost is marginal. However, the computational load for performing simulation with GAP increases with the number of data points in the training set because the kernels of every pair of training points should be included in the model. Therefore, one needs to reduce the number of reference structures in the training set with advanced techniques such as CUR matrix decomposition methods [59].

Throughout this study, we use NNP as a machine learning potential to describe the

potential energy surface of DFT. In this Chapter, we describe the model, descriptor, training techniques for NNP.

2.2.1 Model

NNP adopts a feed forward neural network to describe the structure-property relation. The schematic illustration of a neural network is shown in Figure 2.1. Mathematically, this is formulated as:

$$x_k^{i+1} = \sigma\left(\sum_j^{N^j} x_j^i w_{jk}^i + b_k^i\right), \tag{2.17}$$

where x_j^i indicates the *j*th node of the *i*th layer, and σ is the activation function which gives the non-linearity of the model. The fitting parameters are w_{jk}^i and b_k^i where the former is the weight parameter connecting the *j*th node of the *i*th layer and *k*th node of the *i*+1th layer, and the latter is a bias added to the *i*th layer. For activation functions, sigmoid function, hyperbolic tangent function, and rectifying linear unit (ReLU) functions [60] are usually employed. In this dissertation, we mainly adopt the hyperbolic tangent function as the activation function:

$$\sigma(x) = \frac{\mathrm{e}^{-x} - \mathrm{e}^{x}}{\mathrm{e}^{-x} + \mathrm{e}^{x}}. \tag{2.18}$$

The activation function is not used for connecting the last hidden layer and the output layer. Instead, they are linearly connected.

When constructing the NNP models, one can simply set the atomic coordinates as an input layer and total energy as an output [39]. However, in this case, the length of the input layer changes with the number of atoms. Thus this model is only applicable to the system that consists of the same number of atoms as the reference structures in the training set. To resolve this problem, Behler-Parrinello type NNP (BP-NNP) is introduced [38]. The key idea of this model is to decompose the total energy of the system (E_{tot}) into the sum of atomic energies (E_{at}) of individual atoms in the system

as follows:

$$E_{\mathrm{tot}} = \sum_{i=1}^{N_{\mathrm{at}}} E_{\mathrm{at}}(\mathbf{G}_i),$$

(2.19)

where *i* is the atomic index, N_{at} is the number of atoms in the system and the $\mathbf{G}_i$ is the input descriptor that captures the local environment of atom *i*. E_{at} of each atom is calculated by an artificial neural network. The atoms in the same atomic species share the same NNP. The BP-NNP model is schematically illustrated in Figure 2.2.

The atomic forces (**F**) and stress ($\mathbf{S}^{\mathrm{static}}$) can be also obtained by differentiate equation (2.19) with atomic coordinates [61]:

$$F_{i,\alpha} = -\frac{\partial E_{\mathrm{tot}}}{\partial R_{i,\alpha}} = -\sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{\partial E_i}{\partial G_{i,j}} \frac{\partial G_{i,j}}{\partial R_{i,\alpha}},$$

(2.20)

$$S_{i,\alpha}^{\mathrm{static}} = S_{i,\alpha}^{\mathrm{static,rad}} + S_{i,\alpha}^{\mathrm{static,ang}},$$

(2.21)

$$S_{i,\alpha\beta}^{\mathrm{static,rad}} = \sum_{i=1}^{N}\sum_{j=1}^{N} (R_{i,\alpha} - R_{j,\alpha}) F_{j,\beta},$$

(2.22)

$$S_{i,\alpha\beta}^{\mathrm{static,ang}} = \sum_{i=1}^{N}\sum_{j=1}^{N} (R_{i,\alpha} - R_{j,\alpha}) F_{j,\beta} + \sum_{i=1}^{N}\sum_{m=1}^{N} (R_{i,\alpha} - R_{m,\alpha}) F_{m,\beta},$$

(2.23)

where *i* is the atomic index, α and β denotes the *xyz* index, $\sigma^{\mathrm{static,rad}}$ and

$\sigma^{\text{static,ang}}$ indicate the radial and angular parts of the stress components, respectively. The sum of atomic stresses is equal to the total stress of the system.

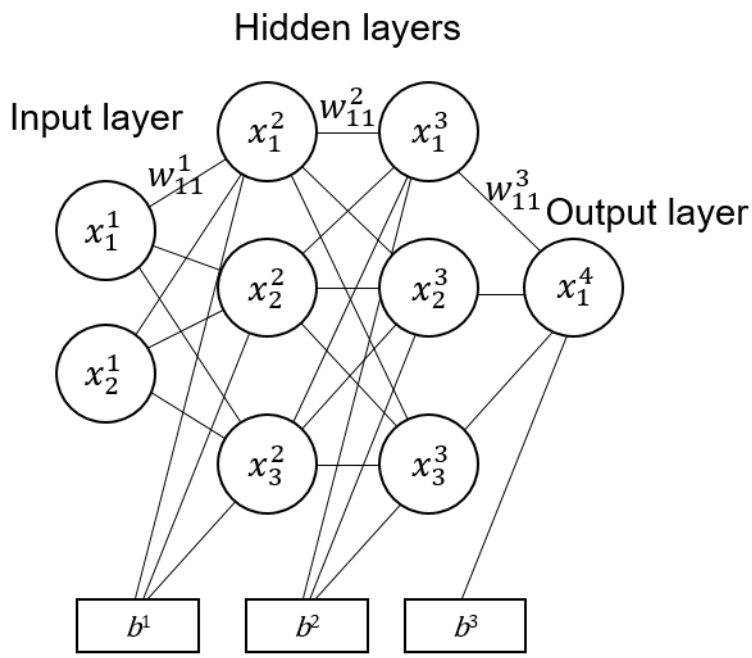

Fig. 2.1: Schematic illustration of artificial neural network.

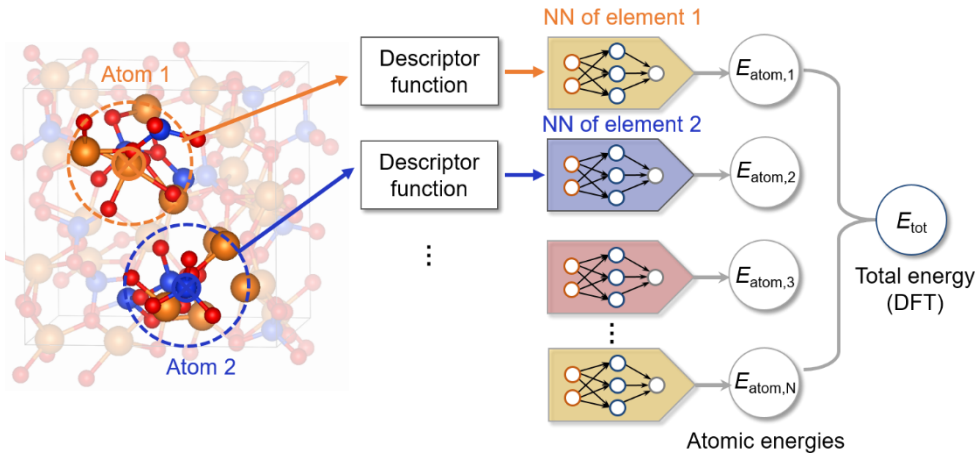

Fig. 2.2: Schematic illustration of neural network potential.

2.2.2 Descriptor

As an input layer, NNPs require descriptors that reflect the local structure information around each atom. Previous studies suggested several descriptors, such as Coulomb matrix [62], bag of bonds (BoB) [63], bispectrum of density [40], smooth-overlap of atomic positions (SOAP) [64], and atom-centered symmetry function (ACSF) [61]. Good descriptors should satisfy the invariance over translation, rotation, and permutation of atoms with the same elements. ACSF and SOAP both satisfy such invariance so that they are widely utilized in diverse chemical systems. In particular, ACSF is known to be well fitted to NNPs.

The Behler-Parinello type of ACSF is the most popular choice with the NNPs [61]. The generally used symmetry functions are radial symmetry function (G^{radial}), and angular symmetry function (G^{angular}) which are written as follows:

$$G_i^{\text{radial}} = \sum_j \mathrm{e}^{-\eta\left(R_{ij}-R_s\right)^2} f_{\mathrm{c}}\left(R_{ij}\right), \tag{2.24}$$

$$G_i^{\text{angular}} = 2^{1-\zeta} \sum_{j,k\neq i} \left(1+\lambda cos\theta_{ijk}\right)^{\zeta} e^{-\eta\left({R_{ij}}^2+{R_{jk}}^2+{R_{ik}}^2\right)} \times f_c\left(R_{ij}\right) f_c\left(R_{jk}\right) f_c(R_{ik}), \tag{2.25}$$

where *i*,*j*, and *k* are the atomic indices, η, ζ, and λ are the hyperparameters, and f_{c} is the cutoff function that is expressed as below:

$$f_{\mathrm{c}}\left(R_{ij}\right) = \begin{cases} 0.5\left[\cos\left(\dfrac{\pi R_{ij}}{R_{\mathrm{c}}}\right)+1\right], & \text{if } R_{ij} < R_{\mathrm{c}} \\ 0, & \text{otherwise} \end{cases} \tag{2.26}$$

Here, R_c is the given cutoff radius. The multiple outputs of G_i^{radial} and G_i^{angular} are obtained within the given set of hyperparameters and are collected as a vector to be used as a descriptor of NNP.

For multi-element systems, symmetry functions are separately calculated for each element pair. For example, in the case of element A in the A-B system, radial symmetry functions of A-A pairs and A-B pairs are calculated. For angular symmetry functions, A-A-A, A-A-B, and A-B-B interaction values are included in the training set. Note that the number of symmetry vectors increases with the number of elements in the system. Therefore, the symmetry function that does not modify the structure when changing the number of elements have been developed [65], but the reliability of this descriptor in various environment should be tested.

2.2.3 Training methods

The training of NNPs is conducted by optimizing the weights and biases. The training process is carried out by minimizing the difference between the errors of target properties predicted by the model. The minimization target is called the loss function (Γ), which is generally defined as follows:

$$\Gamma = \frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{str}}}\left(\frac{E_i^{\mathrm{DFT}}-E_i^{\mathrm{NNP}}}{N_{\mathrm{at},i}}\right)^2 + \frac{\mu}{3\sum_{i=1}^{N_{\mathrm{tr}}}N_{\mathrm{at},i}}\sum_{i=1}^{N_{\mathrm{tr}}}\sum_{j=1}^{N_{\mathrm{at},i}}\left|\mathbf{F}_i^{\mathrm{DFT}}-\mathbf{F}_i^{\mathrm{NNP}}\right|^2 + \frac{\delta}{6N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{str}}}\left(\frac{S_i^{\mathrm{DFT}}-S_i^{\mathrm{NNP}}}{N_{\mathrm{at},i}}\right)^2, \tag{2.27}$$

where E_i^{DFT} (E_i^{NNP}), $\mathbf{F}_i^{\mathrm{DFT}}$ ($\mathbf{F}_i^{\mathrm{NNP}}$), S_i^{DFT} (S_i^{NNP}) are energy, force, and stress of *i*th atom calculated by DFT (NNP), N_{tr} is the number of structures in the training set, $N_{\mathrm{at},i}$ is the number of atoms in the *i*th structure, and μ and δ are the hyperparameters that determine the importance of force and stress in the training process, respectively. Generally, the initial weights and biases of NNPs are set randomly in a normal distribution small standard deviation. The gradient of the weights ($\partial\Gamma/\partial w$) is updated by the backpropagation method. In a simple gradient descent algorithm, each weight is updated by follows:

$$w \rightarrow w - \alpha\frac{\partial\Gamma}{\partial w}, \tag{2.28}$$

where α is the learning rate. α is one of the most important hyperparameters in training artificial neural network [66]: if the learning rate is too small, the

convergence of Γ would be very slow, and if the learning rate is too high, Γ would remain high or even diverge.

To reduce the computational cost for training NNPs, the batch training method is widely used. This method trains over only a few reference data points in each iteration instead of using a full training set. The batch is chosen every iteration. About the optimizer, advanced techniques such as Adam [67], AdaGrad [68], and L-BFGS [69] have been suggested in addition to the gradient descent algorithm. Throughout this dissertation, we choose Adam optimizer with batch training method where the batch size is set to 10-15.

In training a neural network, a tfraction of data points are used as training (which is called training set), and the others are used for validation of the potential (which is called validation set). One significant challenge in neural network training is the overfitting problem which refers to the situation where the validation loss function increases while the loss function of the training set consistently decreases. This often happens when the absolute values of weights become too large [66]. One way to prevent overfitting is to use the regularization method. This method adds an additional term in the loss function that represents the dispersion of the weights. Using this, the minimization of loss function prevents the weight distribution becomes complex. One popular choice is the L2-regularization which adds the sum of the square of the weights in the loss function.

2.3 Crystal structure prediction

Crystal structure prediction (CSP) aims to find the global minimum in the configuration-energy space. Two methods are typically used, the data-mining of the known prototypes [15,16] and heuristic approaches such as random structure generation [20,21], genetic algorithm [28], particle swarm optimization [26,27], simulated annealing [23], minima hopping [24], basin hopping [25] and metadynamics [22]. While the goal of both methods is the same (finding the ground-state crystal structure), the advantages and disadvantages of these methods are distinct. Data mining methods examine the correlation between composition information and structure prototypes. The method is fast because this only requires the probability calculations on a machine-learning model following a few DFT calculations. However, this method cannot predict the crystal structures with new prototypes that do not exist in the chosen database. On the other hand, heuristic approaches directly navigate the configuration space by modifying the given structures. DFT calculations are utilized for local relaxations and evaluations of the free energies. However, this requires a large computation time. Therefore, several techniques, such as topological random generation [70] and antiseed scheme, are proposed to accelerate navigating the configuration space in heuristic approaches. In this Chapter, we review the technical details of the widely used CSP methods, and also discuss the key algorithms that have been developed recently.

2.3.1 Data-mining approaches

The first data-mining method is suggested by Fischer et al. in ref. [15] A machine learning model was trained to represent the correlations between crystal structures in different compositions in the experimental database. The authors first define the representation of entire crystal systems in the database by vector $\boldsymbol{X}$, which is defined as:

$$\boldsymbol{X} = \left(x_{E_1}, x_{E_2}, \dots, x_{E_n}, x_{c_1}, x_{c_2}, \dots, x_{c_n}\right), \tag{2.29}$$

where c_i represents the composition, x_{c_i} represents one crystal structure prototype of which composition c_i can have (e.g., if c_i is AB_2C_4, x_{c_i} can have values such as spinel), and x_{E_i} indicates the system's constituents (e.g., E_i = Ag, Cu, Na, etc). The **e** vector is additionally defined to represent the information on the existing crystal structures of the target system. For instance, the Ag-Zr system can be represented as

$$\mathbf{e} = (x_{E_1} = \text{Ag},\ x_{E_2} = \text{Mg},\ x_{\text{Ag}} = \text{FCC},\ x_{\text{Ag}_3\text{Mg}} = \text{ZrCl}_3\text{ type},\ x_{\text{AgMg}} = \text{CsCl type},\ x_{\text{Mg}} = \text{Mg type}). \tag{2.30}$$

To predict the crystal structure of **e** in undiscovered composition c_u, one needs to evaluate $P(x_{c_u}|\mathbf{e})$ for all possible crystal structures in x_{c_u} (namely, $P(\boldsymbol{X})$). For example, in order to predict the crystal structure of $AgMg_3$, one needs to evaluate all possible $P(x_{c_u}|\mathbf{e})$, such as $P(Cu_3P\text{ type}|\mathbf{e})$, $P(BiF_3\text{ type}|\mathbf{e})$, $P(IrAl_3\text{ type}|\mathbf{e})$, etc, and choose the candidate structures with high probability. $P(\boldsymbol{X})$ can be expanded as follows [71]:

$$P(\boldsymbol{X}) = \prod_i P(x_i) \prod_{j<k} g^2(x_j, x_k) \prod_{l<m<n} g^2(x_l, x_m, x_n) \dots,$$

(2.31)

where g^2 and g^3 denote the cumulant functions which represent the correlation among pairs and triplets of structures, respectively. Generally, this extension is terminated for triplets and higher-order, and can be written as:

$$\mathrm{P}(\boldsymbol{X}) = \frac{1}{Z}\prod_i \mathrm{P}(x_i)\prod_{j<k} g^2(x_j, x_k),$$

(2.32)

where Z is the normalization factor. Morita identified the cumulant functions as follows [71]:

$$g^2(x_j, x_k) = \frac{\mathrm{P}(x_i, x_j)}{\mathrm{P}(x_i)\mathrm{P}(x_j)}.$$

(2.33)

Therefore, the parameters $\{\mathrm{P}(x_i, x_j)\}$ and $\{\mathrm{P}(x_i)\}$ are essential in calculating 2.27. These probabilities are learned from experimental database using Bayesian estimation method.

Hautier et al. devised a similar but different machine learning model for predicting crystal structures [16]. This method examines similarities to whether structures containing two elements have the same prototype. It provides the probability of substitution of one element into other elements when the composition is given. In this method, the compound consisting of *n* different ions are represented as follows:

$$\mathbf{X} = (X_1, X_2, \ldots, X_n), \tag{2.34}$$

where variable X_j are defined in Ω, which denotes the set of existing ion species:

$$\Omega = \{\mathrm{Fe}^{2+}, \mathrm{Fe}^{3+}, \mathrm{La}^{3+}, \ldots\}. \tag{2.35}$$

Then, one needs to know $p_n(\mathbf{X}, \mathbf{X}')$ which is the probability of how likely ion substitution from $\mathbf{X}$ to $\mathbf{X}'$. This is expressed as follows:

$$p_n(\mathbf{X}, \mathbf{X}') = p_n(X_1, X_2, \dots, X_n, X'_1, X'_2, \dots, X'_n). \tag{2.36}$$

$p_n(\mathbf{X}, \mathbf{X}')$ can be a guide to determine how likely the composition $\boldsymbol{X}'$ can have the crystal structure of $\mathbf{X}$.

The probability function is multidimensional so that can not be evaluated directly and only can be approximated. In the literature, this is defined using binary indicator *f*, which is called a feature function. In this case, *f* is defined as follows:

$$f_k^{a.b}(\mathbf{X}, \mathbf{X}') = \begin{cases} 1, & \text{if } X_k = a \ \ and \ \ X'_k = b \\ 0, & \text{otherwise} \end{cases} \tag{2.37}$$

where *a* and *b* denotes the ion index. Then, $p(\mathbf{X}, \mathbf{X}')$ can be approximated as follows:

$$p(\mathbf{X}, \mathbf{X}') \approx \frac{e^{\sum_i \lambda_i f_i(\mathbf{X}, \mathbf{X}')}}{Z}, \tag{2.38}$$

where *i* denotes the possible set of (*a*,*b*,*k*), *Z* is the partition function, and λ's are the fitting parameters. These values can be determined over crystal information in the experimental database.

Using the above methods, high-throughput searches on missing ternary oxides [72] and nitrides [17] were performed. The later findings successfully lead to the experimental realization as stated in the same literature. However, this method is restricted to finding materials with known structure prototypes and fails to identify the global minimum when the prototype information is not known in advance [73].

2.3.2 Heuristic approaches

In this Chapter, we briefly overview the most popular five global optimization methods of CSP.

Random structure sampling

Random structure sampling generates numerous structures randomly and finds the lowest energy structures which is famously implemented in AIRSS [20]. To effectively generate relevant random structures, distance constraints and symmetry constraints are often applied.

Genetic algorithm

Genetic algorithm is the most popular method for crystal structure prediction, which is implemented in several packages such as USPEX [29] and XtalOpt [74]. The schematic illustration of the genetic algorithm is shown in Figure 2.3. In this method, the structures in the pool compete with each other in terms of energy, and only the low-energy structures are survived and are inherited to the next generations. Structures of a fixed proportion are generated randomly and the others are generated from mutations such as crossover, lattice mutation and permutation. With crossover (or heredity) operator, two structures are cut in half and merged to a new structure. The spatially reasonable parts of the structures are selected using order-fitness correlation [31]. The lattice mutation alters the lattice vectors of a structure, and the permutation swaps two atoms of different atomic species. In addition, softmutation algorithm is implemented in USPEX which modify the structure along the direction of the phonons that are approximately estimated from the simple spring model [29].

Particle swarm optimization

Particle swarm optimization (PSO) aims to find the global minimum by iteratively improving the candidate solution by sharing information between the particles [30].

Here, the particles are defined as the positions in the searching space. In the problem of CSP, the individual crystal structures in the pool correspond to the particles. In specific, the particles' positions (x^k) and velocities (v^k) in search space are as follows:

$$x_i^{(k+1)} \leftarrow x_i^{(k)} + v_i^{(k+1)}, \tag{2.39}$$

$$v_i^{(k+1)} \leftarrow v_i^{(k)} + \mathrm{r}_1(p_i - x_i^{(k)}) + \mathrm{r}_2(g - x_i^{(k)}), \tag{2.40}$$

where *k* is the iteration step *i* is the particle index, p_i is the position of the solution that *i*th particle found, and g is the best-known solution. r_1 and r_2 are the hyperparameters that determine the significance of the experiences of particle itself, and the particle with the best solution.

Metadynamics

Metadynamics is the sampling technique that adds the repulsive potential to the already sampled region [75]. (The definition of sampled regions is defined by the collective variables.) Therefore, the molecular dynamics simulation tends to find untrained regions with metadynamics and avoid being trapped in the local minima. Thus this method can search wide configuration space. Usually, this method is used to sample the reaction pathways of which activation barriers are high so that is difficult to be captured by molecular dynamics with finite temperature [76]. In ref. [22] the method is used to discover the high-pressure phases of materials using the components of lattice vectors as collective variables. In addition, the metadynamics method is combined with an evolutionary algorithm to find the crystal structures in ambient conditions [77]. However, this method is challenging to be used as a general CSP method in defining the general collective variables.

Minima hopping

Minima hopping is the global minimization technique that explores the configurational space by avoiding revisiting already visited regions [27]. This algorithm is based on two principles. The first one is the built-in feedback mechanism which recognizes the previously visited regions and prevents revisiting them. The second one is the Bell-Evans-Polanyi principle which states that the free-energy difference has a linear correlation with the reaction barrier. This is not strictly followed in all cases but works in an average manner. The minima hopping algorithm finds crystal structure as following steps: First, the algorithm starts from the current local minima on the configurational space and tries escape step using a short molecular dynamics (MD) simulation in a soft direction. The MD simulations are terminated as soon as potential energy maxima have been crossed certain criteria. Then, a local geometry relaxation is performed. The algorithm determines whether to take the new minimum based on the energy difference. If the energy difference does not exceed the given threshold, the same step is repeated. If the new structure is accepted, another escape is tried at the new local minima.

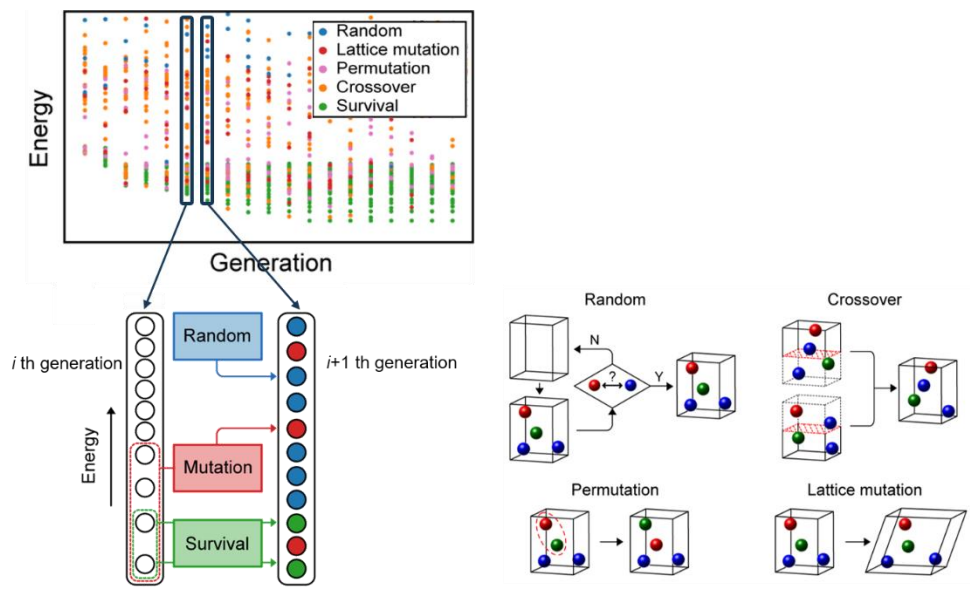

Fig. 2.3: Schematic illustration of the evolutionary algorithm-based crystal structure prediction

2.3.3 Local optimization and energy evaluation

In all heuristic methods, structural relaxations are accompanied by global optimizations after generating the structure. This corresponds to the optimization of the minimum of the near basin in the potential energy surface. This part takes the longest time in the whole CSP algorithm because the cost of DFT calculations is high. The typically used algorithms are the conjugate-gradient (CG) method, quasi-Newton algorithm, and damped molecular dynamics. On the other hand, in the case of CSP, the structural relaxation mostly starts from the randomly generated structures of which chemical orders largely deviate from those of the local minima, accompanying large structural change during relaxation. Therefore, this generally takes much longer relaxation steps compared to the simple ones which marginally changes the local orders. Because the structure relaxations are performed by DFT, the total computational cost is significant. To mitigate this problem, Oganov et al. suggested a multistage strategy, where relaxation proceeds in 3-5 stages [29]. At the initial stage, the accuracy of DFT settings is loosely set, and then the precision of DFT calculations increases with proceeding the rest of the stages. For instance, one can set the **k**-point spacing to 0.2 $Å^{-1}$ at the initial stage, and consistently decrease the value so that the final value is 0.08 $Å^{-1}$. This can reduce the total computation time by a few times.

2.3.4 Structure similarity

The heuristic CSP approaches generate numerous structures during the algorithm. If the identical structures are included in the structure pool, the search efficiency of the algorithm can be lowered. To evaluate whether the structures are the same or not, similarity metrics should be required. Previous studies suggested several metrics such as bond order parameters [78], landscape statics [79–81], disconnectivity graph [82], etc. However, these works are not developed to be used in crystal structure prediction, but for specific purposes such as analyzing kinetics and physical properties of clusters and disordered structures. Thus, it is unclear that these descriptors can judge the similarity of the structures with various local orders generated during the structure search. In this section, we introduce two widely used metrics: fingerprint function and bond characterization matrix.

The fingerprint function is suggested by Oganov et al. [83] and implemented in the USPEX code [29]. The fingerprint matrix (**F**) is a $N \times N$ matrix where N is the number of atoms. The component of **F** for atomic type pairs AB (F_{AB}) is defined as follows:

$$F_{\mathrm{AB}}(R) = \sum_{\mathrm{A}_i,\mathrm{cell}} \sum_{\mathrm{B}_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 \frac{N_{\mathrm{A}} N_{\mathrm{B}}}{V} \Delta} - 1 = g_{\mathrm{AB}}(R) - 1, \tag{2.41}$$

where i is the atomic index of type A, and j is the index of B atoms within the cutoff distance. δ is the delta-function, R_{ij} is the distance between ith and jth atoms, N_{A} (N_{B}) is the number of atom A (B) in the unit cell, and Δ is the width of the bins for F_{AB}. g_{AB} is the partial radial function between atomic species A and B. The final fingerprint of structure is defined as the concatenate of weighted elements of **F**. The weight for the AB components of **F** is defined as:

$$w_{\mathrm{AB}} = \sqrt{\frac{N_{\mathrm{A}} N_{\mathrm{B}}}{\sum_{\mathrm{cell}} N_{\mathrm{A}} N_{\mathrm{B}}}}.$$

(2.42)

The cosine distance are used to measure the distance between structure *I* and *J*:

$$D_{\mathrm{cosine}} = \frac{1}{2}\left(1 - \frac{\sum_{AB} \sum_{k} F_{I,\mathrm{AB}}(k) F_{J,\mathrm{AB}}(k) {w_{\mathrm{AB}}}^2}{|F_{I,\mathrm{AB}}(k) w_{\mathrm{AB}}| \cdot |F_{I,\mathrm{AB}}(k) w_{\mathrm{AB}}|}\right).$$

(2.43)

The bond characterization matrix is developed by Wang et al. and implemented in the CALYPSO code [30]. This metric is a modification of the bond-orientational order metrics (Q_{l}) suggested by Seinhardt et al. [78], where the spherical harmonic and exponential functions are employed to identify the bond angles and lengths, respectively. Specifically, it is defined as:

$$\overline{Q}_{lm}^{\delta_{AB}} = \frac{1}{N_{\delta_{\alpha\beta}}} \sum_{i \in A, j \in B} \mathrm{e}^{-\alpha(r_{ij} - b_{AB})} Y_{lm}(\theta_{ij}, \phi_{ij}),$$

(2.44)

where *A* and *B* are the indices for atomic types, *i* and *j* are indices for individual atoms, $\delta_{\alpha\beta}$, and $N_{\delta_{\alpha\beta}}$ indicate bond type and the number of bond types, respectively, r_{ij} is the interatomic distance of atom *i* and *j*, and *Y* is the spherical harmonics. To make the metric rotationally invariant, the modified metric is defined as follows:

$$Q_l^{\delta_{AB}} = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left|\overline{Q}_{lm}^{\delta_{AB}}\right|^2},$$

(2.45)

where *l* values of 0, 2, 4, 6, 8, 10 were used as a type of bonds in the reference.

Finally, the similarity distance between structure *u* and *v* (D_{uv}) is defined as follows:

$$D_{uv} = \sqrt{\frac{1}{N_{\text{type}}} \sum_{\delta_{AB}} \sum_{l} \left(Q_l^{\delta_{AB,u}} - Q_l^{\delta_{AB,v}} \right)^2}. \tag{2.46}$$

where N_{type} denotes the number of bond types.

2.3.5 Advanced techniques on genetic algorithm

As mentioned above, global optimization takes significant computation time due to the large cost of DFT. Therefore, several techniques have been developed to reduce the number of generations for finding the global minimum. Some important additional techniques used with the genetic algorithm are introduced in this subsection.

Symmetry-based random structure generation

Most of CSP programs employ the symmetry-based random structure generation algorithm [20,29,30,84]. The algorithm first randomly selects the space group and then selects the lattice vectors in accordance with the chosen space group. Then, Wyckoff sites are randomly selected to put atoms in the cell. Then, the structure is checked whether it satisfies the distance constraints. If the structure does not satisfy the constraints, it selects other Wyckoff sites. If this iteration exceeds the given number, then the space group is selected again and repeated the process until the structure satisfies the constraints. As a result, the symmetric structure having a randomly chosen space group is generated. This method is proven to be more effective for CSP than just a simple random structure generation method only using distance constraints [20].

Topology-based random structure generation

To further improve the efficiency of CSP, Bushlanov et al. developed the other random structure generation method from the given topology [70]. The topology is defined by the connectivity of the atoms in the material. In this method, one topology is selected from the ToposPro database [85], and then the structure having this topology is generated. In principle, an infinite number of structures can be generated from one topology. This method is tested to be more efficient in finding the low-energy structures [70]. However, we note that this method is not completely *ab initio*,

because this uses the structure information of materials in the database.

Increasing the diversity

To search the wide configurational space of materials in a short generation, it is important to maximize the diversity of the structure pool during the evolution. In this sense, it would be lower the efficiency of the algorithm if the already-sampled structures survive too long generations. To discourage the already sampled structures from surviving to the next generation, the penalty is added to the survived structures which relate to the similarities of the structure compared to the other previously sampled structures. This potential is name antiseed (*A*) [31] which is defined as:

$$A = \sum_{a} W_a \exp\left(-\frac{D_{ia}^2}{2\sigma_a^2}\right), \tag{2.47}$$

where a indicates the index for structures that are obtained in the previous generations, D_{ia} is the structural distance (see Chapter 2.3.4) between ith and ath structure, and parameter σ_a and W_a is set to proportional to the fitness variance. This antiseed weight is updated every generation. On the other hand, the parameter setting can be very sensitive, so multiple trials over the various set of parameters may be needed.

Multi-objective optimization

The general purpose of CSP is to find the ground-state structures. On the other hand, metastable structures such as diamonds can also have good target properties. To find these materials, multi-objective optimization techniques are employed in evolutionary CSP [86]. In this method, two fitness's are optimized at the same time. The survival of the structure is not determined by energy alone, but structures at the Pareto front of two fitness's (such as energy and band gap) all survive into the next

generation.

Variable-composition method

Most structure searches aim to find the global minimum in the fixed composition. However, the optimal composition is not known in advance in most cases. The variable-composition method is developed to search the various compositions at the same time [33]. This algorithm starts with the initial random structures with various compositions and does not fix the composition during the structure generation by crossover. The fitness corresponds to the hull energy (energy above the convex hull). This can be especially effective for metals and high-pressure phases [33] where various compositions are allowed [87]. However, this may not be favorable for searching semiconducting materials at ambient conditions because the oxidation number is very sensitive to the stability of the phase in this case.

Chapter 3

CSP with machine learning potentials

3.1 Training machine learning potentials

When developing machine learning potentials (MLPs), constructing reference structures for the training set plays the most important role, because MLPs only describe potential energy surfaces from the configurations included in the training set. Therefore, the configurations that may appear during a target simulation must be sampled in advance. This poses a distinct challenge when training MLP for crystal structure prediction (CSP), where the target material is unknown before making predictions. To address this issue, previous researches suggested methods to sample training set by random sampling or evolutionary search in an on-the-fly manner [41–43]. These methods were tested on unary materials such as boron clusters, boron bulk allotropes, high-pressure sodium, and carbon allotropes. However, these methods require ~1000 DFT relaxations which may be inefficient for complex ternary or higher (multinary, henceforth) compounds. In addition, the errors of the MLPs are larger than 40 meV/atom, which were not accurate enough to identify fine energy differences among metastable structures near the global minimum.

Here, we propose a method to construct MLP as a surrogate model of DFT. Our main idea is to use DFT molecular dynamics trajectories of disordered structures

(amorphous and liquid) as a training set to construct MLPs. DFT melt-quench-annealing simulations can be started from scratch, eliminating the need for prior knowledge of chemical information of the target materials. In addition, these trajectories include diverse local orders that may appear during the CSP procedure. Note that this idea was first suggested by Deringer et al. [88]. However, the previous study has tested this idea only on carbon systems that are expected to have the simpler potential energy surface than complex multinary compounds. Therefore, the question remains whether MLPs trained by this method can correctly reproduce potential energy surfaces of more complicated systems.

In this Chapter, we aim to develop the method to train MLPs for multinary systems using disordered structures. We use Behler-Parinello type neural network potentials (NNPs) [38] but the present method can be also used with other types of MLPs, such as Gaussian approximation potentials [40] and moment tensor potentials [55]. The method is tested on four multinary materials, Mg_2SiO_4, $LiAlCl_4$, Ba_2AgSi_3, and $InTe_2O_5F$, of which crystal information is presented in the ICSD database [8]. To test the quality of NNPs, we compare DFT and NNP energies for experimental and metastable structures that appear during DFT-based CSP algorithms. Impressively, NNP consistently rank the energies of experimental structures lower than those of metastable structures in all cases. Using the developed NNPs, we also conduct CSP using evolutionary algorithm program. This identifies the experimental structure of $LiAlCl_4$. On the other hand, the method fails to find the ground states for other materials, but only finds metastable structures within an energy range of 10-41 meV/atom compared to the ground states.

3.1.1 Melt-quench-annealing simulation

We perform a molecular dynamics (MD) simulation of a melt-quench-annealing procedure using DFT calculations for each material to generate a training set for an NNP. All DFT calculations in the present work are performed by using Vienna Ab-initio Simulation Package (VASP) code [89]. We adopt the Perdew-Burke-Ernzerhof (PBE) version of generalized gradient approximation for exchange-correlation functional [90]. The Γ-point is used for **k**-point sampling for all materials, and cutoff energies of Mg_2SiO_4, $LiAlCl_4$, Ba_2AgSi_3, and $InTe_2O_5F$ are set to 450, 300, 250, 500 eV, respectively. These settings are chosen by convergence tests on the superheated structures with energy, maximum atomic force, and pressure limits are set to 20 meV/atom, 0.3 eV/Å, and 10 kbar, respectively. The time step for MD simulations is set to 2 fs.

First, we perform a superheating simulation of 5 ps at 4000 K for a randomly generated structure of ~100 atoms with the same stoichiometry as each target compound. We determine the *ad hoc* melting temperature (T_m) as the lowest temperature at which the mean square displacements of all atomic types are linear with time. T_m of Mg_2SiO_4, $LiAlCl_4$, Ba_2AgSi_3, and $InTe_2O_5F$ are determined to be 3500, 1500, 1500, and 2000 K, respectively. Then the volume of the cell is determined so that the external pressure is zero at T_m. The trajectory of the liquid structure is generated by the MD simulation for 20 ps at T_m, and the structure is quenched at a rate of 100 K/ps to 300 K. We also perform the annealing simulation at 500 K for 15 ps to further sample low-energy amorphous structures. The reference structures for training set are selected every 10 steps.

3.1.2 Training NNP

To train NNP, we use SIMPLE-NN code [91]. We employ a Behler-Parinello type NNP which provides the DFT energy as the sum of the atomic energies predicted by artificial neural networks [38]. We use 132-30-30-1 (212-30-30-1) architectures of NNP for ternary (quaternary) compounds with atom-centered symmetry functions [61] as input layers. The cutoff distance of the symmetry functions is set to 6 Å. The symmetry function vectors are transformed by principal component analysis and then whitened in order to avoid using highly correlated symmetry functions. In addition, all the components of symmetry functions are normalized [92]. The loss function of the training is set to the sum of energy, force, and stress errors and the L2 regularization term. One-tenth of the reference data is randomly chosen and used as validation structures. The training is conducted with ADAM optimization method [67] with a batch size of 10-15, until the root mean square errors (RMSEs) of the validation set become smaller than 10 meV/atom, 0.2 eV/Å (except for $InTe_2O_5F$: 0.4 eV/Å), and 10 kbar for the energies, forces, and stresses, respectively. We also check that the overfitting does not exist during the training. The LAMMPS package is employed for NNP calculations [93].

3.1.3 Evaluation of the quality of NNP

To test the quality of NNP, we choose three ternary and one quaternary compounds from the ICSD, Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$ (see Figure 3.1). These materials contains diverse structural orders: Ba_2AgSi_3 has a layered structure formed by Si_2Ag, with Ba atoms intercalated between them, Mg_2SiO_4 consists of Mg-O octahedron and Si-O tetrahedron and contains both edge- and corner-sharing of them, $LiAlCl_4$ consists of Li-Cl and Al-Cl tetrahedrons where they are all connected by corner-sharing, and $InTe_2O_5F$ is a multi-anion compound with corner-shared In-O-F and Te-O polyhedrons. In addition, their space groups are different from each other (*Fddd*, *Pnma*, *Pmn*2_1, and *C*222_1, for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively). We emphasize that the materials with low symmetries, not simple compounds with high symmetries (e.g., $SrTiO_3$), are chosen to stress-test our method. In addition, these materials have distinct physical properties and are therefore being applied to the various applications: Ba_2AgSi_3 is a member of Ba-Ag-Si system which has been investigated for potential high-T_c superconductors [94]. Mg_2SiO_4 have a high fracture toughness so that it is considered as a material for bioceramic implants [95]. The ICSD database also includes three other metastable phases of Mg_2SiO_4. $LiAlCl_4$ is a Li superionic conductor under aliovalent doping [96]. Two similar-energy structures of $LiAlCl_4$ exist in the ICSD (*P*2_1/*c* and *Pmn*2_1), and we find that *Pmn*2_1 is slightly more stable from DFT calculations within PBE level. Finally, $InTe_2O_5F$ has been studied as a potential nonlinear optical material due to the noncentrosymmetric structure [97]. The theoretical bandgaps calculated within the HSE06 functional [50] are 0.25, 6.70, 7.17, and 4.71 eV for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively.

To investigate the accuracy of the developed NNPs as an energy evaluator in CSP, we compare NNP and DFT energies of metastable structures. Here, the metastable structures are generated by the CSP algorithm using evolutionary algorithm and DFT

calculations, implemented in the Universal Structure Predictor: Evolutionary Xtallography (USPEX) package [29]. The formula unit (Z) of each ground-state experimental structure is used for each materials. The population size is set to about twice the number of atoms in the unit cell, which corresponds to 20-60. We use random seeding, and mutation operators to generate structures. For the former, we use either symmetry-based random structure generation [31] or topological structure generation methods [70]. For mutation operators, we use crossover (also called heredity), permutation, lattice mutation and soft mutation. The initial ratio of operators are set to 16.6%, 16.6%, 41.6%, 8.3%, 8.3%, and 8.3% for symmetry-based random structure generator, topological structure generator, crossover, permutation, lattice mutation and soft mutation, respectively. The proportions of these operators are set automatically in every generations to diversify the structures in the pool. All structures in the pool are relaxed until atomic forces and total stress is less than 0.1 eV/Å and 20 kbar, respectively, or the number of relaxation steps becomes 400. Note that we used multistage relaxation method described in Chapter 2.3.3. The weighted partial radial distribution functions (pRDF) [83] are used as structure similarity metric of the structures. We use antiseed option, which add repulsive Gaussian energies for sampled structures that prevents the global optimization trapping into the local minima [31].

We gather all structures that emerged during the whole evolutionary searches (10-20 generations) and use them in the benchmark test of NNP. On average, 274 metastable structures are sampled. The energies relative to that of the stable phase (ΔE_{tot}) are less than 500 meV/atom for all materials. The lowest ΔE_{tot}'s are 46.5, 28.2, 1.9, and 33.2 meV/atom for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively, which means that the evolutionary algorithm did not find the reference structures.

NNP energies are evaluated within the fixed geometries obtained from DFT relaxations. Figure 3.2 shows the correlation between ΔE_{tot}'s evaluated by DFT

($\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$) and those obtained by NNP ($\Delta E_{\mathrm{tot}}^{\mathrm{NNP}}$). It is noticeable that NNP and DFT energies are highly correlated especially in the low-energy regions ($\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$ < 200 meV/atom). The Pearson coefficients in this region are 0.769, 0.864, 0.977, and 0.962 for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively. This is impressive considering the fact that the NNPs were only trained over disordered structures and none of the metastable crystal structures used for the tests were not included. This means that the melt-quench-annealing trajectories sample diverse structural motifs contained in low-energy metastable structures. In addition, each NNP consistently ranks the most stable experimental structure to be the most stable among all structures appeared in Figure 3.2. We also observe the same ordering when the structures are fully relaxed using NNP. This indicates that NNP can act as a suitable surrogate model of DFT in the CSP algorithms.

In Figure 3.2, RMSEs of $\Delta E_{\mathrm{tot}}^{\mathrm{NNP}}$'s with respect to $\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$'s are 27.4, 29.6, 10.7, and 63.7 meV/atom for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively, in the low-energy region ($\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$ < 200 meV/atom). These errors are larger than the training errors because local motifs in disordered structures may not exactly equal with those in the crystalline phases. Nevertheless, the trained NNPs can still serve as a surrogate model in CSP because final candidates in certain energy window are re-evaluated by DFT calculations at the final stage of CSP (discussed in section 3.2.4).

It can be seen in Figure 3.2 that the NNP errors in high-energy region ($\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$ > 200 meV/atom) are larger than those of lower-energy structures. This might be attributed from the fact that the structural motifs in high-energy structures were not sufficiently sampled in disordered structures: the disordered structures are generated entropy-driven, so the high-energy phases have low chances to be sampled. It is also noticeable that the systematic upward deviation exists in Figure 3.2 rather than the downward deviation. This is because the structures are first relaxed by DFT, and then evaluated by NNP. Therefore, there exist slight deviations from the equilibrium positions of NNP. We also observed the opposite trends when DFT energies are

evaluated by NNP-relaxed structures.

To further verify whether the training set encompass the configurational space of metastable structures, we plot distributions of symmetry function (**G**) vectors of the training set and those of metastable structures of Ba_2AgSi_3. In Figure 3.3, the **G** vectors are aligned by PCA, and the distribution over principal axis is shown. It can be seen that most **G** points from metastable crystal structures and experimental structures lie within those from the training set. Other materials also show similar trends to that of Ba_2AgSi_3. The only exception is Al in $LiAlCl_4$, where the **G** components of Al atoms from high-energy metastable structures largely deviate from the training structures. However, the **G** components of low-energy structures well fitted in the training set. Therefore, it is explicitly verified that the local orders of metastable structures generated by the evolutionary algorithm and the experimental structures are well included in the configuration space of the training set.

Even though the training set contains overall local motifs similar to the target structures, the exact local motifs in the low-energy metastable structures may not be included in the liquid and amorphous structures. This can be seen in Figure 3.2a that the large error (~100 meV/atom) occurs near the ground-state structure of Ba_2AgSi_3. This is because any hexagonal Si_4Ag_2 in the experimental structures does not exist in the training set. This might be attributed from the high quenching speed and small simulation cell. This problem can be resolved by using an iterative training scheme that adds metastable structures to the training set and refines NNPs (see section 3.2.1).

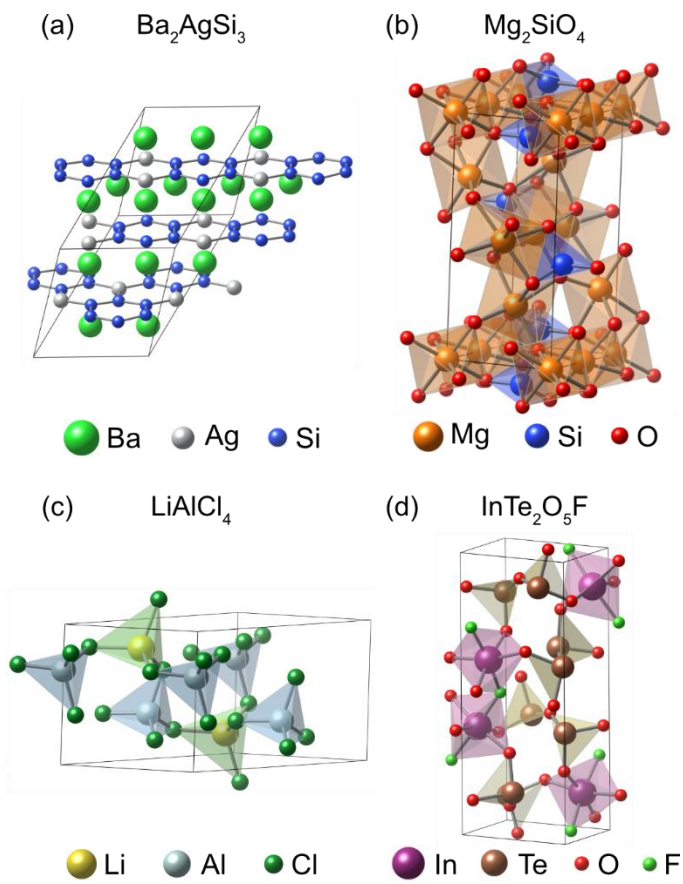

Fig. 3.1: The crystal structures of (a) Ba_2AgSi_3, (b) Mg_2SiO_4, (c) $LiAlCl_4$, and (d) $InTe_2O_5F$.

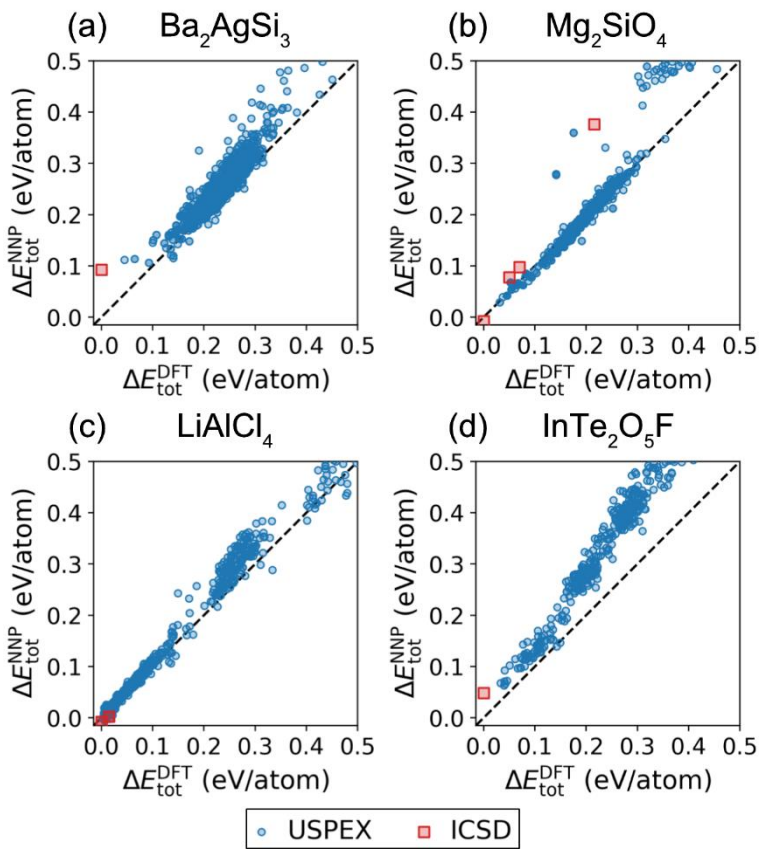

Fig. 3.2: Correlation between DFT and NNP energies. For both $\Delta E_{\mathrm{tot}}^{\mathrm{NNP}}$ and $\Delta E_{\mathrm{tot}}^{\mathrm{DFT}}$, the reference energy is the DFT energy of the experimental phase. Experimental structures are plotted as red squares. (a) Ba_2AgSi_3, (b) Mg_2SiO_4, (c) $LiAlCl_4$, and (d) $InTe_2O_5F$

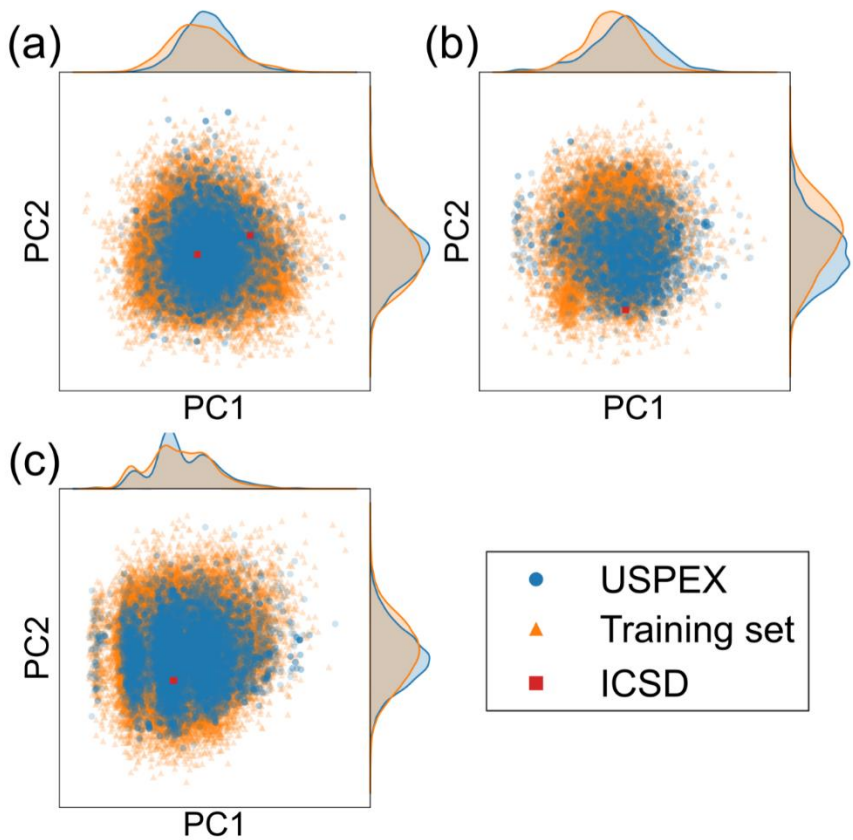

Fig. 3.3: The distribution of **G** of Ba_2AgSi_3 projected onto the first two principal-component axes (PC1 and PC2). The distribution of (a) Ba (b) Ag, and (c) Si atoms.

3.1.4 Structure searching with NNP

In the last Chapter, we confirm that NNPs trained over disordered structures can be a hi-fidelity surrogate model for CSP. Therefore, we perform CSP with constructed NNPs for the four test materials. We interface USPEX [29] with LAMMPS [93] code for the implementation of NNPs in structural relaxations and energy evaluations. The computational parameters for USPEX are set the same as those in Chapter 3.1.3, except this time we set the number of generations to 120. This is longer than 10-20 generations that were available within DFT-USPEX but the computational time is much shorter due to the much faster speed of NNP. At the end of the 120 generations, we to relax 10-20 lowest-energy candidate structures with DFT utilizing AMP^2 code [98]. The relaxation is performed until all the atomic forces and the stress components are less than 0.02 eV/Å and 4 kbar.

Despite the longer generations compared to the conventional settings in previous studies (mostly for unary or binary materials) [31,99,100], USPEX fails to identify the most stable experimental structures except for $LiAlCl_4$. This may represents the complicated configuration spaces of multinary compounds. The structures with the lowest ΔE_{tot} values are shown in Figure 3.4 and the corresponding energies are 41.2, 27.7, 0, and 11.5 meV/atom for Ba_2AgSi_3, Mg_2SiO_4, $LiAlCl_4$, and $InTe_2O_5F$, respectively. While the structures of three candidate structures are similar to the ground states, there are local orders are distinct from them: final candidate of Ba_2AgSi_3 (Figure 3.4a) consists of only Si-Ag rings but does not contains exact Si_4Ag_2 and Si_6 rings in the experimental structure (Figure 3.1a). Mg_2SiO_4 found by USPEX with NNP (Figure 3.2a) only has tetrahedral Mg and Si atoms while they are corner-sharing with each other by O atoms. However, this is largely deviate from the most stable experimental structure of Mg_2SiO_4 which contains octahedral Mg atoms with mixed corner- and edge-sharing O atoms. Nevertheless, the NNP energy of this structure is close to the experimental structure, which implies that the ground-state structure can be found in longer generations. In the case of $InTe_2O_5F$, the

experimental and the predicted structures have similar structural motifs with threefold Te and twofold O. However, the major differences are observed for local orders around In: the In atoms in the experimental structure (Figure 3.1d) are corner-shared with other In atoms by F atoms, resulting in twofold F atoms. However, the predicted structure (Figure 3.4d) contains edge-sharing singly coordinated F. Even so, the ΔE_{tot} of $InTe_2O_5F$ is only 11.5 meV/atom, which implies that it has a rugged energy landscape where the structural similarity is weakly correlated with the energy differences.

So far, we set *Z* as those of the most stable experimental structures. However, this information is not known when searching unknown materials for the practical purpose. Therefore, one has to try various *Z* values to find the correct ground state. In this sense, NNP is expected to be much more advantageous than DFT in searching with various *Z* due to the order-*N* scaling. (DFT scales with order-N^3.)

About the computational cost, the main load originates from constructing the training set by melt-quench-annealing simulation by DFT calculations. It took about a couple day to generate training set and a half day for training NNPs using 4-CPU (8-CPU) clusters for ternary (quaternary) materials. The Intel Xeon Phi 7250, 1.4 GHz (68 cores per CPU) is used in this study. CPU times for structural relaxations in CSP were 2,072 (69,594) and 11,532 (244,646) s/CPU for $LiAlCl_4$ and $InTe_2O_5F$, respectively. It took 1 day for NNP to finish 120 generations of USPEX calculations and 3-4 days for DFT to finish 10-20 generations.

As mentioned above, CSP of complicated multinary compounds may require generations much longer than 100. In principle, NNP can reach this generation because the calculation speed of the structural relaxation and the energy evaluation using NNP is ~1000 times faster than DFT. However, the speed of USPEX using NNP for 120 generations is only 30 times faster than USPEX using DFT. This means that the fast speed of NNP is not fully reflected in the CSP algorithm with USPEX. We speculate this by three reasons: First, the structure generations and pRDF

calculations of generated structures are done sequentially. The whole process takes about few minutes. The longest part is the random structure generation using topology information [70]. This load is not taken seriously when using DFT-based CSP because this is marginal compared to the calculation time of DFT calculations. However, when using NNP, the structural relaxation takes only about few minutes, so the structure generations take similar time to the structural relaxations. This load would be much more reduced if it is done parallel. The second reason is that the parallelization of the structure relaxations in USPEX is done by submitting multiple queue jobs. When using NNP, the computation time for one job of structural relaxation mostly takes less than a minute (the longest job can take few minutes), so this puts excessive loads on the main CPU for uploading and unloading multiple jobs in a short time. This often results in the systematic time delays and weakens the stability of the cluster server. Third, we find that the rest of the post-processing time increases linearly with the number of generations. We cannot find the exact reason for this because the main code of USPEX is not open in public. We speculate that it is attributed to the comparison of pRDFs of new 4structures with all the previous structures for eliminating duplicate structures and calculating antiseed weights.

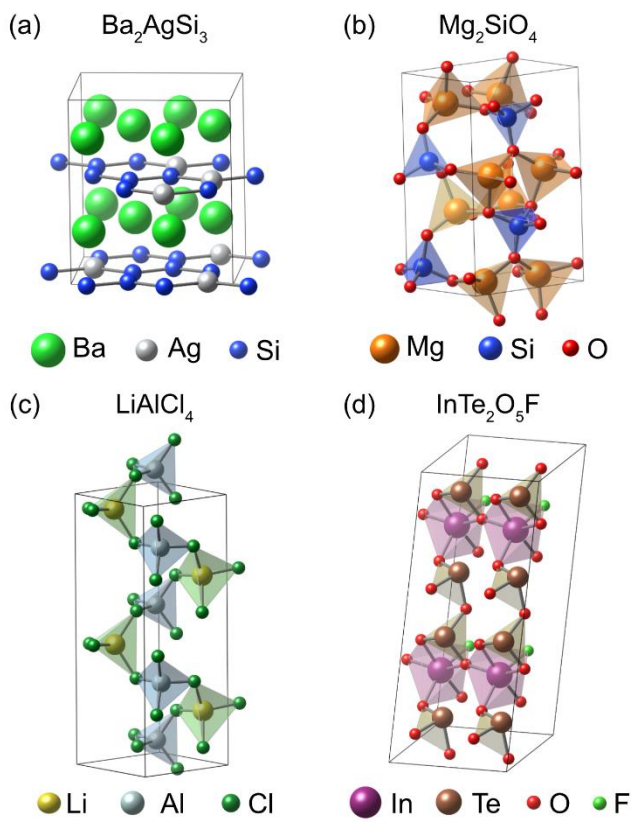

Fig. 3.4: The most stable structures of (a) Ba_2AgSi_3, (b) Mg_2SiO_4, (c) $LiAlCl_4$, and (d) $InTe_2O_5F$ found by NNP-USPEX.

3.2 Developing and optimizing CSP algorithm

In this Chapter, we develop the NNP-based CSP algorithm that can address the problems mentioned in Chapter 3.1.4. First, we optimize the training procedure of NNPs to further increase the accuracy of NNP and reduce the computational cost of constructing training set. Then, we optimize the global optimization algorithm to fully harnessing the speed of NNP. Finally, we demonstrate the parallelization scheme that are used in this study. We note that modifying other CSP code is not an option because the most of popular CSP codes such as USPEX [29] and CALYPSO [30] are not open in public.

3.2.1 Optimization of training procedure

As mentioned in Chapter 3.1.4, the computational cost to generate the training set takes the most significant amount of time during the whole process. To reduce the time cost, we reduce the number of atoms in the unit cell from 100 to 80 in MD simulations. This reduces the computational costs in half. In addition, we reduce the melting time from 20 ps to 16 ps, increase quenching speed from 100 K/ps to 200 K/ps, and reduce annealing time from 15 ps to 5 ps to further save the computational cost. The computational time for training of NNPs and structural relaxations with NNPs can be further reduced by changing the cutoff radius of the **G** vectors. We reduce the cutoff radius of the angular parts of **G** vectors from 6 to 4.5. This reduces the training time and structure relaxation time by about half. Note that the mentioned changes may sacrifice the accuracy of the NNP. Nevertheless, we demonstrate that the quality of NNPs are good enough in most cases for CSP which will be discussed in the section 3.3.

The above-mentioned settings are used for the calculations in Chapter 3.2. On the other hand, we confirm that the further increase in computational speed is possible by further loosening the settings with the additional tests. We reduce the number of atoms in the unitcell to ~64. In addition, melting time is set to 10 ps, and quenching is done only for 1500 K (e.g., 2500 K $\rightarrow$ 1000 K). We use this setting for the calculations for applications of CSP in Chapter 4.

As mentioned in Chapter 3.1.3, the large deviation of NNP energies from DFT energies (~100 meV/atom) is found near the ground-state structure of Ba_2AgSi_3 in Figure 3.2a. This is because the Si and Si-Ag hexagonal rings are not exactly included in the training set. The metastable structures sampled from the NNP-based CSP can be utilized as an additional training set to update NNPs. To test this idea, we add 22 lowest-energy candidate structures of Ba_2AgSi_3 to the training set and construct the new NNP. In specific, we included the DFT relaxation trajectories of candidate materials and their distorted geometries due to the elastic deformation in

the training set. Figure 3.5 shows the correlation of DFT and NNP energies using the trained potential. (We use the same test structures that were employed to generate Figure 3.2a.) Compared to Figure 3.2a, much more accurate correlations between NNP and DFT energies are obtained, which is a result of configurations including Si-Ag rings to the training set. In specific, the Pearson coefficient in the low-energy region ($\Delta E_{\mathrm{tot}}^{\mathrm{NNP}}$) increases from 0.769 to 0.828 after refining the potential. This significant improvement in NNP can boost identifying the global minimum in CSP.

This advancement in Ba_2AgSi_3 can occur generally in other materials as well because the exact local orders in the training set do not exactly match the local orders in the training set. We construct a systematic approach to iteratively refine a training set which is presented in Figure 3.6. With the constructed NNP from melt-quench-trajectories, we proceed 50 generations of genetic algorithm and collect 10 structures. 5 lowest-energy structures and 5 structures with the lowest antiseed weights are sampled within 100 meV/atom. The antiseed weight represents the degree of distinction of a given structure from the other structures in the pool. Here, the antiseed weight of *i*th structure is defined as follows:

$$A_i = \sum_a \exp(-\frac{{d_{ia}}^2}{2\sigma^2}), \tag{3.1}$$

where the summation runs over structures in the low-energy pool (bottom 200 meV/atom), d_{ia} is the similarity distance between the *a*th and *i*th structures measured by pRDF, and σ is a Gaussian width. Unlike the original scheme [31], we did not consider all the structures that emerged in the previous generations but only include the structures in the low-energy pool in the current generation. The relaxation trajectories of the collected structures are included in the training set to refine NNP. The structure relaxation is performed until the maximum atomic forces in the cell and the stress components are less than 0.1 eV $Å^{-1}$ and 20 kbar, respectively. We

find that the initial cell structure sometimes changes significantly during the relaxation due to the imperfect accuracy of NNPs. Therefore, the **k**-point spacing is monitored every 10 relaxation steps and adapted if the spacing becomes different from the original **k**-point spacing. Then, we proceed CSP and refine NNPs in the same way at 100, 200, and 400 generations, but remove structures already sampled from the previous generations. During evolutionary algorithm, we use symmetric random generation [31], crossover algorithm, permutation and lattice mutation as structure generators.

We find that this iterative scheme is effective in training high-accuracy NNPs in many cases (further discussed in the Chapter 3.3). In particular, the crossover algorithm is found effective in sampling diverse structural motifs that may be absent in the original training set. One intriguing example is the refinement of the NNP of BKS_2. The experimental structure of BKS_2 is shown in Figure 3.7a: it has a B-S hexagonal ring structure connected with K atoms. Before perform iterative training, the ground state structure is not predicted as the lowest-energy structure as shown in Figure 3.8a. On the other hand, the crossover algorithm samples the metastable structure in Figure 3.7b, which contains a ring structure of B-S. After training this configuration, the correlation of NNP energies and DFT energies of metastable structures and the experimental structure is significantly improved (Figure 3.8).

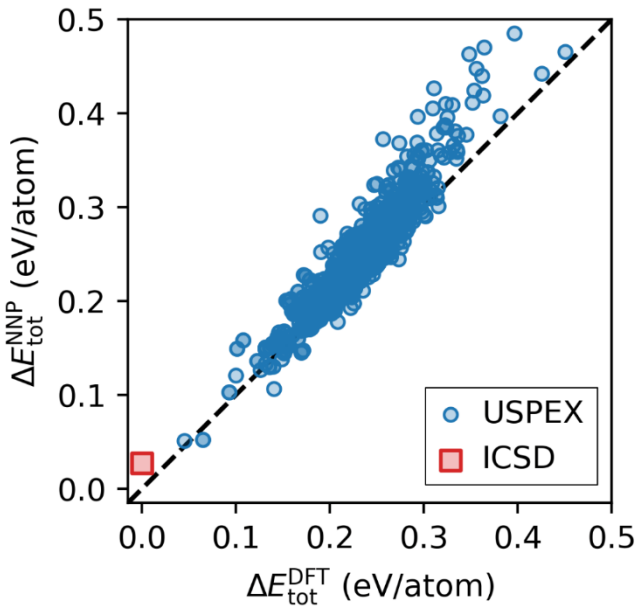

Fig. 3.5: Energy correlation between DFT and refined NNP for Ba_2AgSi_3 with structures in Fig. 3.2a after collecting the training set with structures from USPEX-NNP.

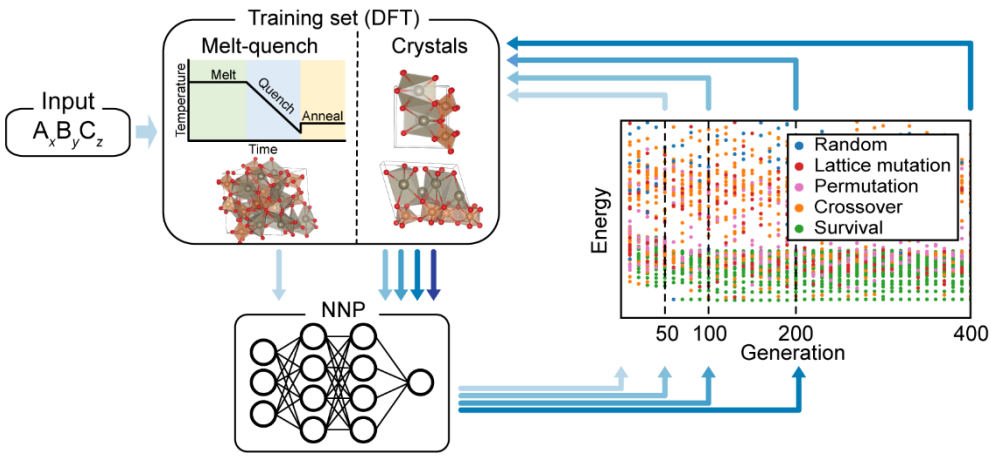

Fig. 3.6: Iterative training scheme of NNP.

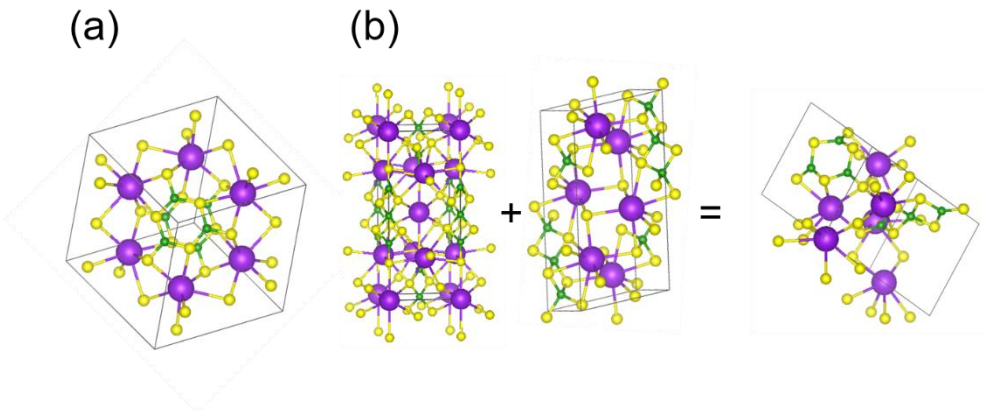

Fig. 3.7: (a) The experimental structure of BKS_2. (b) The sampling of the six-fold ring by crossover algorithm.

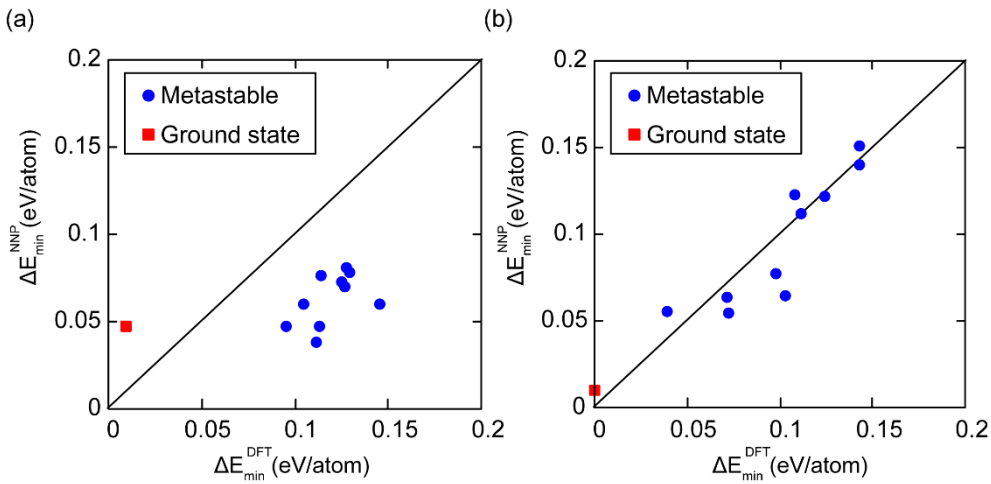

Fig. 3.8: Energy correlation of (a) before and (b) after refining NNP.

3.2.2 Optimization of global optimization

As discussed in Chapter 3.1.4, the previous CSP codes are only optimized for DFT-based CSP methods that need to find ground states in short generations due to the large computation time of DFT calculations. Therefore, this takes large efforts in pre-processing and post-processing in order to effectively explore the configurational space. On the other hand, we failed to find the global minimum of Mg_2SiO_4, Ba_2AgSi_3, $InTe_2O_5F$ with USPEX [29] using NNP. This implies that the current CSP codes are not well optimized for searching the complicated multinary compounds with NNPs which is expected to take large generations. Here, we develop an efficient CSP algorithm that can fully harness the speed and accuracy of NNPs.

We first develop the evolutionary algorithm similar to the conventional methods [28] as shown in Figure 2.3. In every generation, the fraction of structures in the pool are newly generated by random structure generation, and the others are inherited from the low-energy structures (200 meV/atom for refining stages and 100 meV/atom for main CSPs) in the previous generations by crossover, permutation and lattice mutation. The structures within 100 meV/atom (50 meV/atom) energy window further survives from the last generation during refining (main CSP) stage. For random structure generation, we use symmetry-based structure generation method implemented in RandSpg code [84], which is similar to the random structure generation methods in other CSP codes such as USPEX [29] and CALYPSO [30]. This method creates random structures by randomly selecting space groups, lattice vectors, and Wyckoff positions. In the case of the crossover algorithm, we use atomic energies to rationally select slabs of inheriting structures and translation vectors when emerging structures. We use similar permutation and lattice mutation algorithms that are implemented in the DFT-based CSP code [29].

The crossover algorithm tends to sample diverse local motifs, which is useful in training hi-fidelity NNPs (also discussed in Chapter 3.1.4). On the other hand, we

find that the crossover algorithm is not effective in finding the global minimum because it consistently generates low-energy structures especially for large unit cell. This does not effectively navigate a rugged energy landscape in multinary compounds as the case of $InTe_2O_5F$ discussed in Chapter 3.1.3. (The effect of structure generation operators for finding global minimum is further discussed in Chapter 3.3.1.) Therefore, when refining NNPs, the ratios of random generation, crossover, permutations, and lattice mutations are set to 30%, 50%, 10%, and 10%, respectively, while during the final CSP to the find global minimum, we eliminate crossover and the ratios are changed to 70%, 20%, and 10% for random structure generation, permutations, and lattice mutations.

Figure 3.9 shows the parallelization scheme of our algorithm. To reduce the pre-process and post-process times, we perform structure generation, structure relaxation, and pRDF calculation in parallel. In addition, the pRDFs are compared only among the low-energy structures not among all structures that emerged during the previous generations. The main code is written in Python. The parallelization is carried out using mpi4py module. Structure generation and relaxation is performed by RandSpg code [84] and LAMMPS code [93], respectively. Both of codes are written in C++, so we bind these code to the main Python code using pybind11 module. The speeds of the algorithms are close to the original versions of C++ code, which is much faster than Python-based codes. Therefore, the total pre-process and post-process times only take few seconds.

Unlike DFT calculations, NNP calculations always run the risk of falling into untrained domain during structural relaxation. To minimize this risk, we use constraints for atomic pairs during the generation of random structures. This highly prevents the generation of unphysical structures as initial structures that largely deviate the training set. The distance constraints are determined for each atomic pair as the minimum distance of the atomic pair that appear in the melt-quench-annealing trajectory. This constraints often extend to 2–3 Å for some pairs, which are much

tighter than simple constraints preventing too short bonds (usually set to 0.7–1 Å). Therefore, this filters out 99% of structures in the structure generation stage. The same distance constraints are used as the cutoff distances for the restrain option during the structure relaxation. (The restrain option is implemented in LAMMPS code.) The restrain option prevents atom pairs to be in the cutoff distance by applying repulsive harmonic force when the two atoms are closer than the cutoff distance.

To further reduce the computation time for structure relaxation, we first relax the structure with the fixed lattice. Then, the structure is inspected whether it violates the distance constraint and the vacuum constraint (6 Å). If it does not violate both constraints, it is further checked whether the energy is lower than 0.5 eV/atom compared to the lowest-energy structure in the previous generation. If the structure satisfy this criteria, relaxation further proceeds to structural relaxations of both atomic coordinates and lattice parameters. If the structure does not satisfy one of these constraints, the relaxation stop after the lattice-fix relaxation and then the energy is read.

Even though we applied distance constraints to minimize the risk, the unphysical structures that are not in the training set can appear during the simulation in any time. Most of these structures have high energies in NNP that are discarded in the next generation, so does not affect the performance of CSP. On the other hand, some unphysical structures with high DFT energies can have the lowest NNP energy among the other structures in the pool even though the structure satisfy the distance constraint. In this case, only unphysical structures survive into the next generation. To prevent this situation, we monitor the DFT energy of the lowest structures every 1000 generations (see Figure 3.10).

We construct a CSP code by integrating the developed algorithms, which is named as SPINNER (Structure Prediction of Inorganic crystals using Neural Network potentials with Evolutionary and Random searches). The summary of the algorithm is presented in Figure 3.11. For an input chemical composition (elements and

stoichiometry), SPINNER conduct a melt-quench-annealing simulation and trains an NNP over the reference structures consists of this disordered structures. Then, NNP is iteratively retrained over ordered structures in low-energy region. In the final CSP proceeding up to 5000 generations, SPINNER gathers low-energy candidate structures (< 50 meV/atom), which are finally sorted after relaxations by DFT calculations.

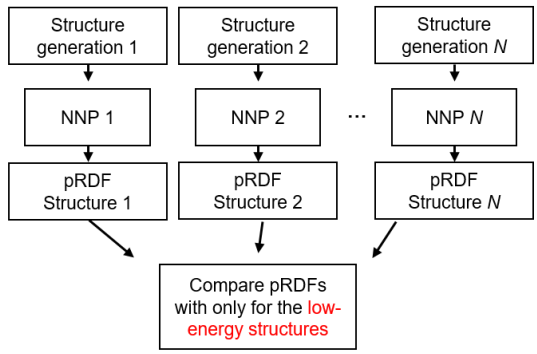

Fig. 3.9: Parallelization scheme of developed code in this dissertation.

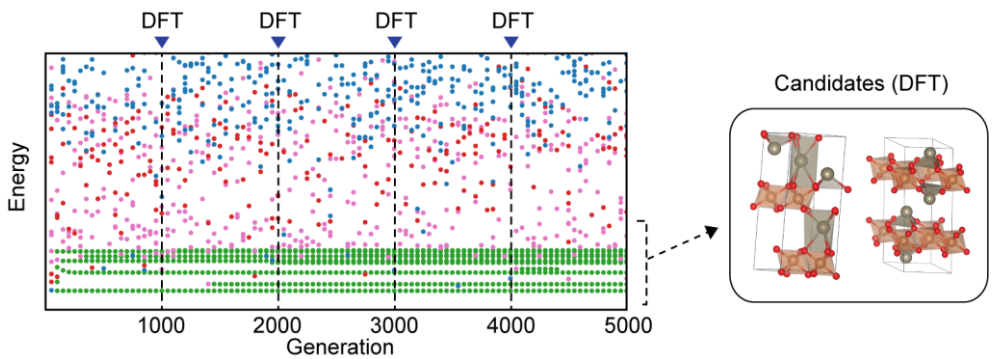

Fig. 3.10: Schematic illustration of main CSP with quality monitoring.

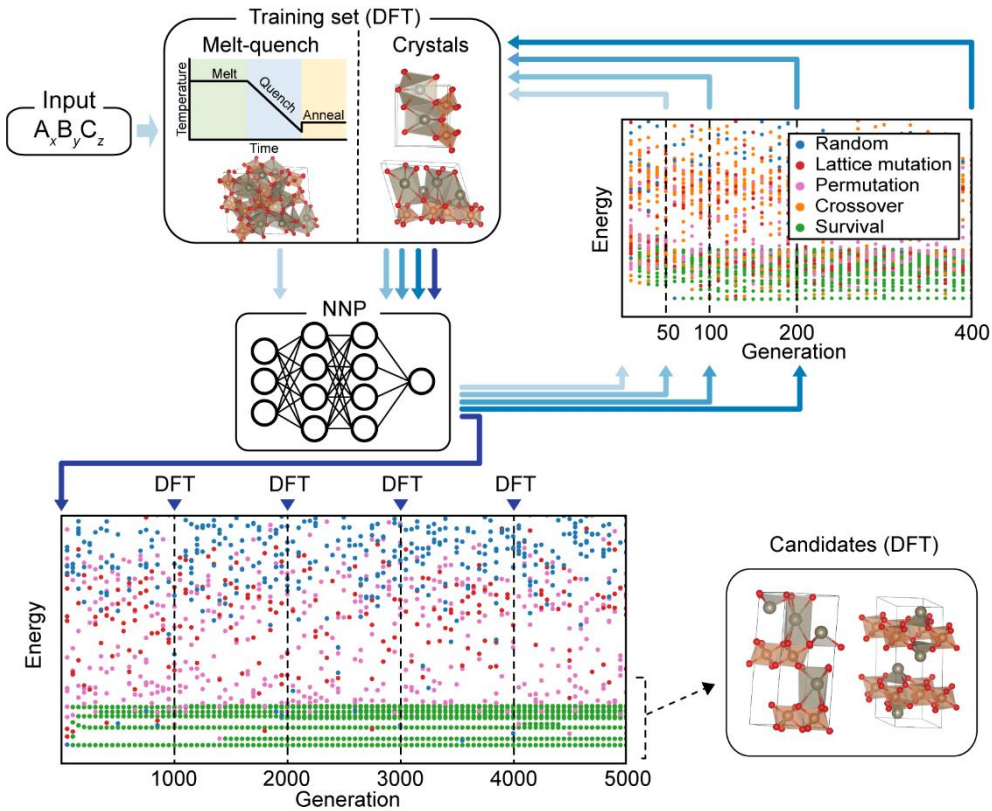

Fig. 3.11: Schematic illustration of the entire algorithm of SPINNER.

3.3 Performance test

In this Chapter, we conduct the performance test of the developed algorithm by two tests: the first one is the blind test over randomly selected experimental structures and the second one is the benchmark test on theoretical structures that are predicted by other CSP methods. Then, we evaluate the computational cost of our algorithm. The results of both tests suggest that the SPINNER can mostly identify the global minimum of the multinary species in a reasonable time scale.

3.3.1 Blind tests on experimental database

In testing search algorithms dealing with NP-hard problems, the global minimum is usually not known. However, the reference structures in CSP is already known because the overall minimum of the potential energy surface for a given thermodynamic and chemical condition is identified by experiment. Therefore, it is possible to estimate the performance of an algorithm by performing blind tests with configurations reported in experimental databases such as ICSD.

In this study, we select ternary compounds as reference structures for the blind test because the database of ternary compounds contains various chemical and structural motifs. First, we randomly select 50 materials based on five criteria: (1) we choose crystals that have high-quality ($R < 0.1$) ordered phases and are stable under ambient conditions. (2) We exclude molecular crystals and only select inorganic crystals. (3) We also did not select compounds including 3d transition metals (V-Zn), lanthanides, and actinides because NNPs cannot delineate magnetic orderings in those materials. (Note that the meaningful achievements are made recently [101,102]). (4) To stress test the algorithm, we constrain the unit (Z) to be greater than 4. (The minimum number of atoms in the cell are at least 12.) (5) Then, we force that at least one crystal is selected from 32 crystallographic point groups. If the ICSD has multiple distinct crystal structures in a composition, the most stable structure within the PBE functional is chosen as the reference structure. We additionally handpick 10 materials that have diverse local motifs and chemistries. The information of 60 selected materials are listed in Table 3.1. Among 60 selected materials, 18 of them are metals, 13 are semiconductors (band gap is between 0 and 2 eV), and 29 are insulators (bandgap is over 2 eV). (The band gaps are calculated with one-shot hybrid functional scheme using AMP^2 code [98].) Most structures have hull energy of 0 in the Materials Project database [103] which means that they are stable under phase separation.

In the blind test, we set Z values as the value of the experimental structures. Figure

3.12a shows the energy difference of the predicted structure by SPINNER and the reference structure (ΔE_{min}) after proceeding with 5000 generations of structure searching. Since the number of the structures in the pool is 24-80, up to half a million structure relaxations are performed for each composition which is enabled by the fast speed of NNP. The color of the data represents the generation at which the minimum energy structure is emerged (N_g). SPINNER correctly identifies reference structures ($\Delta E_{min} = 0$) or lower-energy structures ($\Delta E_{min} < 0$) in 45 compositions out of 60 (75%). Most of them (38 out of 60) are identified within 1000 generations. At the bottom of Figure 3.12a, the unit cells of successfully identified structures are displayed. Among the failed cases, the largest ΔE_{min} occurs for $Sr_2Pt_3In_4$ (36 meV/atom).

In Figure 3.12a, the six materials have negative ΔE_{min}, which means that the structure predicted by SPINNER is more stable than the reference structures in the ICSD database within the PBE functional. We note that in all cases both structures have similar local orders and the difference is only about 4 Å in the distance range. This might attribute to the internal errors in the PBE functional. In fact, the PBE functional has been reported to incorrectly rank the order of metastable phases for some binary materials [52]. In the previous research, it is reported that the SCAN functional partly resolved this issue [51,104]. To verify whether SCAN can correctly rank the energy orders in these 6 cases, we re-relax the structures and evaluate the energies of the predicted structures and the reference structures with the SCAN functional. The empty squares in Figure 3.12a are the ΔE_{min}'s obtained by the SCAN functional. The ΔE_{min} values becomes positive except for LiYSn and $PbOsO_3$. On the other hand, we tested SCAN functional on the 10 opposite cases ($\Delta E_{min} = 0$), the SCAN functional correctly rank the energy orderings of the reference structures to be the lowest. This verifies that the SCAN functional is more accurate than PBE in respect of the energy order. Furthermore, the existence of spin-orbit coupling (SOC) can correct the accuracy of the PBE functional especially for materials including heavy elements. We recalculate the energies of Tl_3PbCl_5 and $PbOsO_3$ with SOC and

find that ΔE_{min} for Tl_3PbCl_5 and $PbOsO_3$ increases to −0.2 and 16 meV/atom, respectively (see sun crosses in Figure 3.12). This demonstrates that the relative stability should be further checked by introducing more sophisticated methods. (Even so, the PBE functional correctly identifies equilibrium structures in many materials including Tl, Pb, and Bi).

The accuracy of NNPs are evaluated by two metrics: the first one is the absolute energy difference (ΔE_0) between DFT and NNP for the reference structures which are relaxed by each method. This demonstrates how well NNP reproduces the structure and energy of the global minimum. The second metric is the averaged energy difference ($\Delta \bar{E}$) of metastable structures within an energy window of 50 meV/atom. This relates to how NNP well describes the potential energy surface near the global minimum. The values of ΔE_0 and $\Delta \bar{E}$ are presented in Table 3.1. The ΔE_0 and $\Delta \bar{E}$ of successful cases ($\Delta E_{\mathrm{min}} \leq 0$) are 12.9 and 11.8 meV/atom, respectively. This means that the potential energy surfaces of NNP and DFT are in good agreement near the global minimum. The remaining errors are partly attributed to the lack of resolution to delineate medium- to long-order correlations beyond the cutoff radii of the descriptors (6 and 4.5 Å for radial and angular parts, respectively). This can be further improved by using more sophisticated descriptors, but this increases the computational cost of the computation.

In contrast to the impressive error of NNPs for the successful cases, ΔE_0 and $\Delta \bar{E}$ for the unsuccessful cases ($\Delta E_{\mathrm{min}} > 0$) are 41.0 and 43.3 meV/atom, respectively, which are approximately four times greater than the successful cases. In particular, $SnGeS_3$ and $Sr_2Pt_3In_4$ showed largest ΔE_0 ($\Delta \bar{E}$) of 75.5 (227.3) and 83.5 (140.0) meV/atom, respectively. Therefore, locating equilibrium structures by NNP would be much more difficult for these materials. The origin of the poor quality of NNPs is not completely resolved yet. (Increasing the cutoff radii of the descriptors was not helpful.) On the other hand, the qualities of NNPs are sufficiently low for some materials in the failed cases. In these materials, it is expected that the ground

minimum will be found in principle by increasing the number of generations over 5000.

We note that $BaGe_2S_5$, Na_3SbO_3, and YPdGe have rather pointed primitive cells while their conventional unit cells have more isotropic with larger *Z*. However, RandSpg [84], the random structure generation code only generates random structures based on the conventional unit cell, which may lower the search efficiency. We further conduct CSP with *Z* equal to that of the conventional cells for these materials. In all cases, SPINNER identifies experimental or theoretically more stable structures (see diamonds in Figure 3.12a).

To analyze how SPINNER effectively finds the global minimum, we analyze the discovery process of the cases for $\Delta E_{\mathrm{min}} \leq 0$ based on two parameters, N_{t} and N_{m}, schematically illustrated in Figure 3.12b. N_{t} is the number of generations from the random structure to the first appearance of the equilibrium structure at N_{g}. N_{m} indicates the number of mutations that takes to find the experimental structure from the initial random structure. The distributions of Nt and Nm are shown in Figure 3.12b. It is intriguing that N_{t} is 0 for almost half of the cases, which means that a minimum-energy structure is obtained directly by relaxing a random structure without any mutations. Even in the cases that take mutations, N_{m} is mostly within 5. This indicates that the relaxation of randomly generated structures is very close to the global minimum. We think there are two reasons to this: first, pair-wise minimum distance constraints filter out ~99% of the structures that are unlikely to be relaxed into physically relevant structures. It also efficiently prohibits atomic configurations from relaxing into high-energy unphysical structures during relaxations. Second, we find that more than half of the initial random structures, evolving into the global minimum within relaxation with NNP, are not relaxed into the same structure when relaxed by DFT but are relaxed into the high-energy metastable structures. This indicates that the potential energy surface at the high-energy regions is significantly different between DFT and NNP. This can be rationalized by follows: NNP

unintentionally smooths the potential energy surface in the high energy region because it is not included in the training set while DFT adaptively forms chemical bonds that can stabilize the initial structure and the relaxation stops at a high-energy metastable structure (illustrated in Figure 3.12c). As a result, the potential energy surface of NNP becomes more efficient than DFT.

We also study whether SPINNER can discover experimental metastable structures other than the one in global minimum. According to ICSD, $TlSbO_3$, $TlGaSe_2$, Na_3PS_4 have ordered metastable compounds with high diffraction qualities ($R < 0.1$) and Z less than or equal to the ground state. These structures lie below the energy range of 50 meV/atom within PBE and are all successfully identified by SPINNER.

To test SPINNER on compounds other than in ternary compositions, we perform CSP on TiO_2, P_3N_5, $NbPd_3$, $Li_{10}GeP_2S_{12}$, and $InGaZnO_4$ whose experimental structures can be extracted from the ICSD. SPINNER successfully finds the ground states within 5000 generations. Among them, TiO_2 is well known for its rich polymorphism [105]. With Z values of 4 and 8, SPINNER identifies all the experimental polymorphs ($C2/m$, anatase, brookite, columbite, rutile) within 50 meV/atom. Even though the test is conducted on the small set of materials, these results support that the SPINNER effectively finds ground states regardless of material complexity. Nevertheless, more complicated quaternary or higher-order materials may generally require longer generations than in this work, which will be the subject of future study.

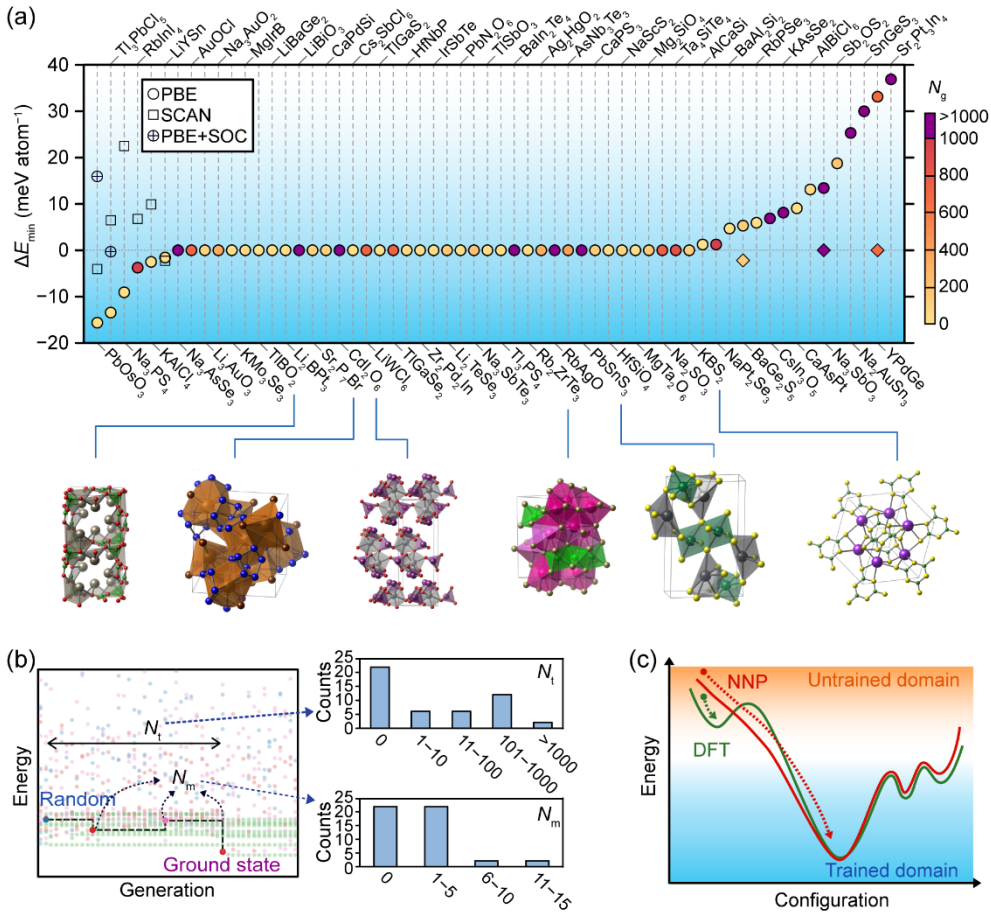

Fig. 3.12: Main results for the blind test on experimental structures. (a) The energy difference between the lowest-energy structures predicted by SPINNER and the reference experimental structure. Some identified crystal structures are shown below for example. (b) Distribution of N_t and N_m. (c) Schematic potential energy surface of NNP and DFT.

Table 3.1: The information of test materials for blind tests. The columns under ICSD are data on the most stable phase in the ICSD. Z and N_{at} denote the numbers of formula units and atoms in the unit cell, respectively. Band gaps (E_g's) are calculated by one-shot hybrid functional calculations [98] and hull energies (E_{hull}'s) are extracted from the Materials Project [103]. We additionally distinguish the energy difference calculated by SCAN functional (†), or by PBE when the spin-orbit coupling is considered (‡). Also, the numbers with the * mark indicate the energy difference between the most stable structure found within ICSD structure conventional cell size and the ICSD structure. The unit for E_{hull}, N_g, ΔE_0, $\Delta\bar{E}$, and ΔE_{min} are meV/atom.

	ICSD						SPINNER			
Formula	ID	Point group	Z	N_{at}	E_g (eV)	E_{hull}	N_g	ΔE_0	$\Delta\bar{E}$	ΔE_{min}
$PbOsO_3$	23444	$m\bar{3}m$	4	20	0	0	5	64.1	20.3	–15.6/ –4.0†/ 16.0‡
Tl_3PbCl_5	1262	4	4	36	6.0	9	11	3.8	5.9	–13.6/ 6.0†/ -0.2‡
Na_3PS_4	72860	$\bar{4}2m$	2	16	3.3	0	1	8.1	16.4	–9.1/ 22.3†
$RbInI_4$	36601	3m	6	36	3.5	0	990	2.9	11.9	–4.0/ 6.4†
$KAlCl_4$	1704	2	4	24	6.8	0	238	2.9	10.1	–2.7/ 9.5†
LiYSn	32041	6mm	8	24	0	0	13	3.3	12.6	–1.7/ –2.0†
Na_3AsSe_3	50491	23	4	28	2.8	0	4004	1.4	7.3	0
AuOCl	8190	$\bar{3}$	6	18	2.3	0	723	22.0	20.1	0
Li_3AuO_3	15113	4/mmm	4	28	3.5	0	80	9.4	11.2	0

Na_3AuO_2	62066	4/mmm	4	24	3.1	0	344	2.5	3.1	0
KMo_3Se_3	603628	6/m	2	14	0	0	6	2.1	4.8	0
MgIrB	409979	622	6	18	0	0	46	1.1	4.3	0
$TlBO_2$	36404	4	8	32	3.8	0	5	0.1	4.7	0
$LiBaGe_2$	162583	mmm	4	16	0	0	17	9.1	13.6	0
Li_2BPt_3	156466	432	4	24	0	0	7	3.9	5.2	0
$LiBiO_3$	82277	mmm	8	40	1.2	0	3032	16.1	5.5	0
Sr_2P_7Br	429306	23	4	40	2.8	0	98	61.6	26.4	0
CaPdSi	69790	2/m	4	12	0	2	90	9.0	19.8	0
CdI_2O_6	1397	222	4	36	4.5	0	3030	9.0	19.8	0
Cs_2SbCl_6	49706	4/mmm	4	36	1.6	0	4	21.7	12.2	0
$LiWCl_6$	409938	3	4	32	0	12	739	8.1	14.4	0
$TlGaS_2$	157537	2/m	8	32	2.4	0	3	0.0	7.1	0
$TlGaSe_2$	1573	m	8	32	2.1	4	737	32.6	28.1	0
HfNbP	75009	mmm	4	12	0	0	32	3.7	2.6	0
Zr_2Pd_2In	107332	4/mmm	4	20	0	135	4	3.7	6.9	0

$IrSbTe$	640967	23	4	12	1.3	0	22	9.9	6.0	0
Li_2TeSe_3	415121	2/m	4	24	0.19	0	161	12.7	10.1	0
PbN_2O_6	174004	$m\bar{3}$	4	36	5.5	0	8	0.2	8.6	0
Na_3SbTe_3	75513	23	4	28	2.0	16	190	14.1	12.3	0
$TlSbO_3$	4123	$\bar{3}m$	4	20	3.3	135	27	5.2	7.5	0
Tl_3PS_4	201062	mmm	4	32	2.7	0	21	10.2	12.4	0
$BaIn_2Te_4$	41168	mmm	2	14	1.6	0	2802	10.1	16.9	0
Rb_2ZrTe_3	410735	2/m	4	24	0.14	0	25	44.4	8.6	0
Ag_2HgO_2	280333	422	4	20	1.3	0	321	22.1	16.3	0
$RbAgO$	40155	4/mmm	4	12	2.7	0	3577	19.8	9.7	0
$AsNb_3Te_3$	79934	6/m	2	14	0	0	408	48.2	35.5	0
$PbSnS_3$	23462	mmm	4	20	1.6	7	1528	35.1	16.0	0
$CaPS_3$	405192	2/m	4	20	4.1	0	131	4.9	11.9	0
$HfSiO_4$	31177	4/mmm	2	12	7.0	0	35	14.9	27.2	0
$NaScS_2$	644971	$\bar{3}m$	1	4	2.7	0	68	5.0	13.7	0
$MgTa_2O_6$	202688	4/mmm	2	18	4.2	0	9	2.4	7.3	0

Mg_2SiO_4	15627	mmm	4	28	6.4	0	152	1.8	3.0	0
Na_2SO_3	31816	$\bar{3}$	2	12	6.4	41	761	3.0	7.0	0
Ta_4SiTe_4	40207	mmm	4	36	0	0	860	11.6	9.6	0
KBS_2	79614	$\bar{3}$m	6	24	3.8	0	113	10.1	17.3	0
AlCaSi	155193	6	6	18	0	0	5	15.3	47.9	1.1
$NaPt_2Se_3$	78788	6mm	4	24	1.6	0	989	2.2	7.6	1.2
$BaAl_2Si_2$	249559	mmm	4	20	0	0	4	40.6	20.7	4.6
$BaGe_2S_5$	66868	m$\bar{3}$m	4/16	32/128	3.2	0	131	43.3	12.7	5.3/−2.3*
$RbPSe_3$	173419	32	6	30	2.0	0	8	57.6	15.3	5.7
$CsIn_3O_5$	23630	mmm	4	36	2.8	0	2017	1.2	6.0	6.8
$KAsSe_2$	65297	1	4	16	2.3	0	4909	40.9	34.2	8.0
CaAsPt	60828	4mm	6	18	0	0	10	7.4	4.0	9.0
$AlBiCl_6$	414261	2/m	4	32	5.0	0	23	14.3	7.6	13.0
Na_3SbO_3	23346	$\bar{4}$3m	4/8	28/56	4.1	0	4958	2.4	8.1	13.3/0.0*
Sb_2OS_2	12120	$\bar{1}$	8	40	1.8	1	172	38.0	16.1	18.6
Na_2AuSn_3	107556	6/mmm	4	24	0	0	2357	36.6	8.3	25.3

$SnGeS_3$	411241	2/m	4	20	1.9	0	4993	75.5	227.3	30.0
YPdGe	391466	mm2	6/12	18/36	0	0	680	21.6	9.9	32.9/0.0*
$Sr_2Pt_3In_4$	410703	$\bar{6}$m2	4	36	0	0	3575	83.5	140.0	36.6

3.3.2 Benchmark test on other CSP methods

To compare the accuracy of SPINNER with other CSP algorithms, we conduct a benchmark test on ternary compounds that were theoretically predicted by either data-mining known prototypes [17,18,106,107] or using DFT-based evolutionary algorithms such as a genetic algorithm [73,87,108] and particle swarm optimization [109]. The materials are listed in Figure 3.13a. As far as we know, none of these compositions has been synthesized except for $KScS_2$, $Sc_2C_3N_6$, and ScPtBi. We perform CSP using SPINER on these compositions for 1000 generations with Z ranging from 2 to 8. The energy difference between the predicted structure and the reference structure in the literature (ΔE_{min}) is shown in the upper part of Figure 3.12a. (The structures extracted from the references are again fully relaxed within the present calculation methods.) SPINNER successfully identifies lower-energy structures in most cases (13 out of 21) and the same structures for the rest. For $Sn_5S_4Cl_2$ and Cd_4SF_6, CSP is performed within the PBEsol functional [110] instead of PBE in ref. [73]. For comparison, we calculate ΔE_{min} with PBEsol and find that the compounds identified in this study are still more energetically favorable than the reference structures by 31 and 4 meV/atom for $Sn_5S_4Cl_2$ and Cd_4SF_6, respectively. We note that $Sn_5S_4Cl_2$, Cd_4SF_6, and $TaCN_3$ have positive hull energies for both reference structures and those identified by SPINNER. Nevertheless, any sign of phase separations is found in the final structures (see Cd_4SF_6, and $TaCN_3$ in Figure 3.13a). This demonstrates that the lower-energy structures discovered by SPINNER are not artifacts of the phase separation. The lower plot of Figure 3.13a represents Z values of the primitive unitcells of the predicted structure by SPINNER (solid squares) in comparison with those in the references (solid circles). SPINNER finds the lowest-energy structure often at a larger Z than that of the reference structures. This indicates that multiple trials of CSPs with diverse Z numbers are important. In this respect, the NNP is much more advantageous compared to DFT owing to the linear scaling of the computation time with respect to the number of atoms, in

contrast to the cubic scaling of DFT.

The structures found by SPINNER usually have similar local structural motifs to reference structures (e.g. $KScS_2$ and $W_4Mo_4B_{15}$ in Figure 3.13b). On the other hand, different local orders can be found in materials such as Li_2TiN_2, Cd_4SF_6, and $TaCN_3$ (see Figure 3.13b). It is intriguing that Cd_4SF_6 and $TaCN_3$ were discovered by DFT-based evolutionary searches. We think that the failure of the previous research in identifying the current low-energy structures might be attributed to smaller *Z* numbers than our calculations and fewer generations. In Figure 3.13a, Na_3OsN_2 has the largest energy difference. Both structures have the same local orders, but the reference structure is significantly distorted. We note that metastable structures can have distinct materials properties from the ground state. For instance, the tetragonal phase of HfO_2 is metastable by 57 meV/atom but has a much higher dielectric constant (70) than that of the monoclinic phase (16) [111]. This stresses the importance of finding the true global minimum in CSP for the reliable prediction of materials properties.

We check the existence of prototypes in the ICSD for the 13 identified structures employing AFLOW-XtalFinder [112]. We find that the structures of 10 of 13 materials do not match any existing prototypes in the ICSD (star-marked in Figure 3.13a). This illustrates that the present prototypes of ternary materials are not sufficiently discovered. It is noticeable that the prototype of $Zr3O_3N_2$ and $Ti_3O_3N_2$ is Ti_3O_5 which is the same as the structure reported in the reference. However, the energies of both structures are lower than those of the structures in the literature. We think that ref. [106] missed the exact ground state because the partial ion exchanges of O and N are not fully considered.

$Sc_2C_3N_6$ has been recently synthesized with the structure predicted by data-mining known prototypes and was also identified by SPINNER [113]. About $KScS_2$, there was an experimental report [114] on the synthesis of the material, but not recognized by ref. [107]. The experimentally found structure is consistent with the present work,

which is slightly more stable than that of ref. [107] by 2.5 meV/atom. The crystal structure of ScPtBi was predicted and identified by the synthesis in ref. [18]. Although it was discovered as a multiphase, SPINNER did not recognize any metastable structure within 50 meV/atom.

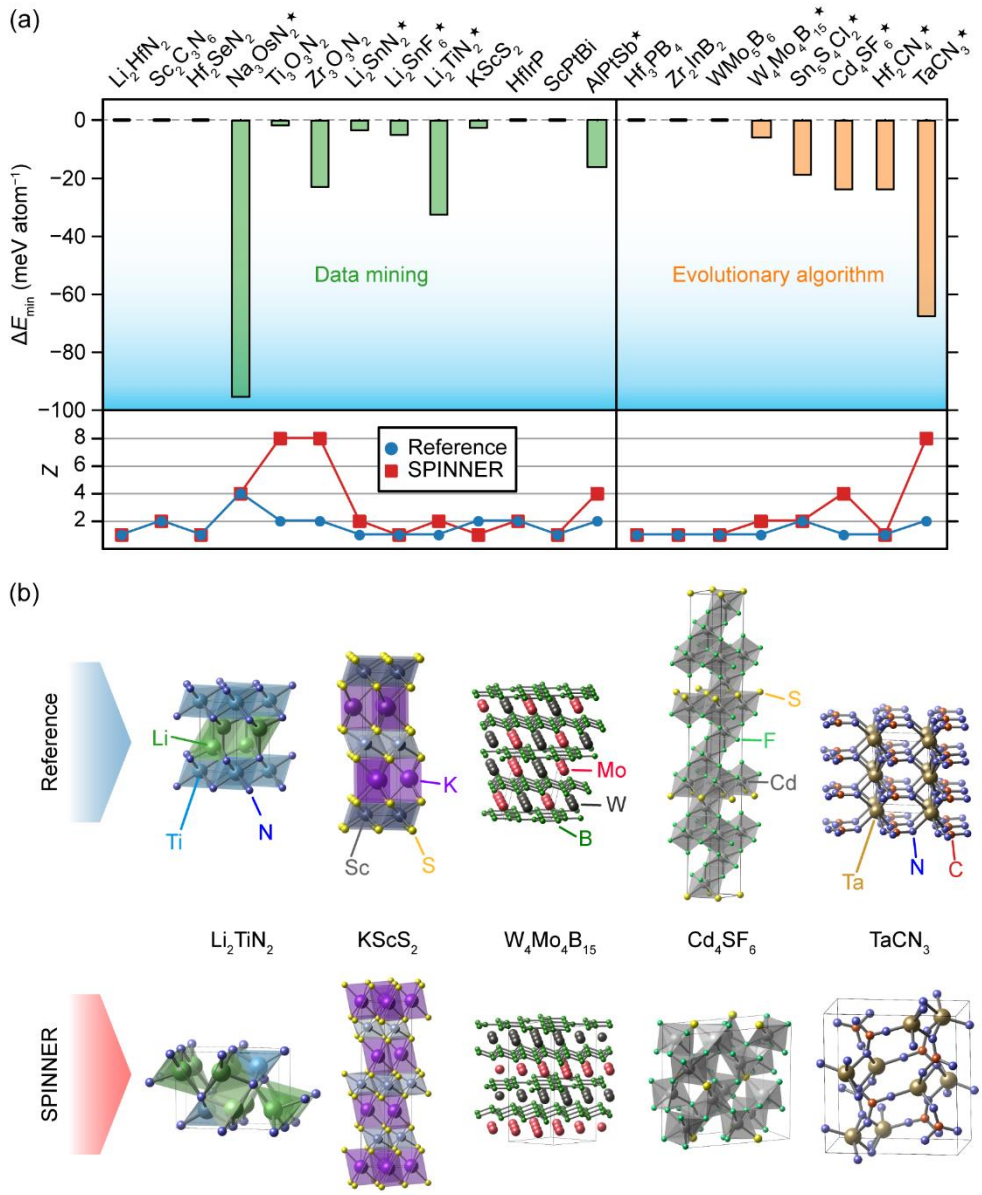

Fig. 3.13: (a) Upper plot shows energy difference between the structure predicted by SPINNER and the structure predicted in literature (ΔE_{min}) by data mining or evolutionary algorithms [17,18,73,87,106–109,115]. The lower part compares the Z between the reference structure and the value at which the lowest energy is found. The compounds marked with stars do not have corresponding structural prototypes in the ICSD. (b) Examples illustrating the structural difference between the crystal structures in references (top) and ones identified by SPINNER (bottom).

3.3.3 Computational cost

About the computational cost, the whole process from the melt-quench annealing molecular dynamics to CSP steps takes 3–5 days on a 36-core node (CPU of Intel® Xeon Platinum 8000-series). For 1000 evolution steps, it takes 3–5 days. On average, the workload of DFT-MD, training NNP, SPINNER, and DFT relaxation of crystals take about 25%, 10%, 60%, and 5%, respectively. The computational time for each part varies widely depending on the materials system. The SPINNER part is largely scalable using many parallel computers. As an example of $LiWCl_6$, it takes 84 hours for 5000 generations with a population size of 64 with $Z = 4$. When tested using identical resources and conditions, a DFT-based evolutionary algorithm [29] only proceeds up to 6 generations under the suggested setting in the manual. In addition, we compare the energy evolution over time using USPEX and SPINNER as shown in Figure 3.14. The energies of the most stable compound in each generation remain larger than 300 meV/atom using USPEX during ~100 CPU hours. The first few generations of SPINNER also remain similar energies compared to USPEX. However, the energy drops quickly over generations, and the reference structure is found around 1000 generations. This only takes ~80 hours including DFT-MD and NNP training time. The estimation of the computational cost demonstrates that it would be possible to construct large databases of as-yet-synthesized materials at a reasonable cost and in a reasonable time scale.

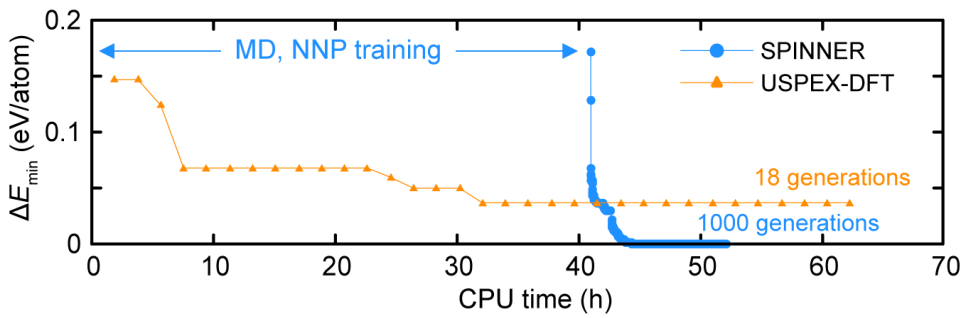

Fig. 3.14: Comparison of the energy evolution over computation time of USPEX and SPINNER. The test material in this Figure is Mg_2SiO_4.

3.4. Transfer learning over compositions

Despite the impressive performance of the SPINNER, the trained NNP is only effective for one composition and may be difficult to be utilized in other compositions because the training set is constructed within a specific stoichiometry. Nevertheless, dynamic fluctuations in the melt-quench-annealing process may extend the local composition range of the training set, which extend the transferability of NNP into other composition. For instance, we conduct CSP for $MgSiO_3$ with the NNP developed for Mg_2SiO_4 and the experimental structure is successfully identified by SPINNER. However, we observe the constant energy shifts between DFT and NNP energies for $MgSiO_3$ because NNP was trained over a single stoichiometry so the atomic energy offsets among elements become arbitrary [92]. To mitigate this problem, we test the transfer learning: the initial NNP trained over disordered structures in the composition of Mg_2SiO_4 are refined over a small number of $MgSiO_3$ crystals which effectively removes the energy offsets. We also verify that the transfer learning well works even for the cases involving valence changes: $MoPdO_5 \rightarrow MoPdO_4$ and $InPbO_3 \rightarrow In_2PbO_4$. This transfer learning scheme is expected to be effective when one tries to discover many possible stoichiometries for unknown materials.

Chapter 4

Applications of CSP

4.1 Synthesizability of missing ternary oxides

Oxides are the most widely studied materials: the frequency of the oxygen in the ICSD is three times higher than that of the most frequent elements (Fe, Si, and S). Indeed, all combinations of binary metal oxides (for which atomic number less than 83) have been synthesized and reported. On the other hand, we find that ~24% of combinations of ternary metal oxides are still missing in the three major experimental databases, ICSD [8], COD [116], and ICDD [117]. Figure 4.1 shows the missing combinations of ternary metal oxides in the three databases. The gray colored combinations exist in the database, and the white colored combinations cannot be found in the database. (We exclude semi-metals here.) In addition, we represent the cases in black when the full phase diagram of the system exists in the FactSage database [118], but no ordered phase exists in the diagram. In this case, the stable ternary metal oxides are not expected to exist at the ambient condition. The atom tags in Figure 4.1 is ordered by the scarcity of the elements. It is noticeable that the number of missing compounds increases with increasing scarcity. Specifically, a large portion of metal oxides including 5d elements are not been identified by experiments yet. On the other hand, many of Be compounds are missing although

the scarcity of the element is not poor. We believe this is due to its toxicity preventing potential researchers from trying to synthesize Be-containing compounds.

Crystal structure prediction can identify whether the missing combinations of oxides are synthesizable (but not synthesized yet) or not-synthesizable. We conduct CSP with SPINNER on missing combinations of ternary metal oxides. For simplicity, we consider the simple composition conditions: the most stable oxidation numbers of elements are employed, the composition is fixed to the 1:1 of pseudobinary materials (e.g., Ag-W-O: $(Ag_2O)_1(WO_3)_1 = Ag_2WO_4$), and two formula units (Z), 3 and 4. Our assumption here is that materials with low hull energies in the 1:1 composition may also have low hull energies in other compositions. To test this idea, we conduct CSP for 7 materials in which any stable phase in the 1:1 composition is not reported, but only in other compositions. We find that all of these materials have low hull energies (< 50meV/atom in PBE and < 40meV/atom in SCAN functional).

Figure 4.2 shows the results for the searching of ternary metal oxides in 1:1 composition. Here, we exclude 3d metals (V-Cu), lanthanides, and actinides due to the magnetic ordering in the materials. The hull energy of the most stable compound in each composition is represented as colors. Among 173 compositions, we find 45 stable compounds. For compositions composed of low scarcity elements (Al-Ga), we only find two compounds (Rb_2MgO_2 and $RbBaO_2$) with negative hull energies. On the other hand, negative-hull-energy materials are much more frequently found in combinations containing scarce elements. Particularly, Au is predicted to form 15 stable compounds with other elements. This trend indicates that most combinations of earth-abundant elements are well studied due to the high economical interest and the most of unreported compositions in this domain are indeed unsynthesizable. On the other hand, the large number of synthesizable combinations with earth-deficient elements indicates that the materials in this domain are not synthesized yet due to the lack of enough effort to synthesize.

As future work, we first plan to complete the search on all missing combinations.

On the other hand, there might be a need to further investigate some compositions with high engineering interests (namely, with low scarcity). We plan to search other composition than 1:1 for combinations where the hull energies in the 1:1 composition are lower than 50 meV/atom (such as Mg-La-O and Zr-Ga-O). The transfer learning technique (see Chapter 3.4) can be used to construct the neural network potential in each composition.

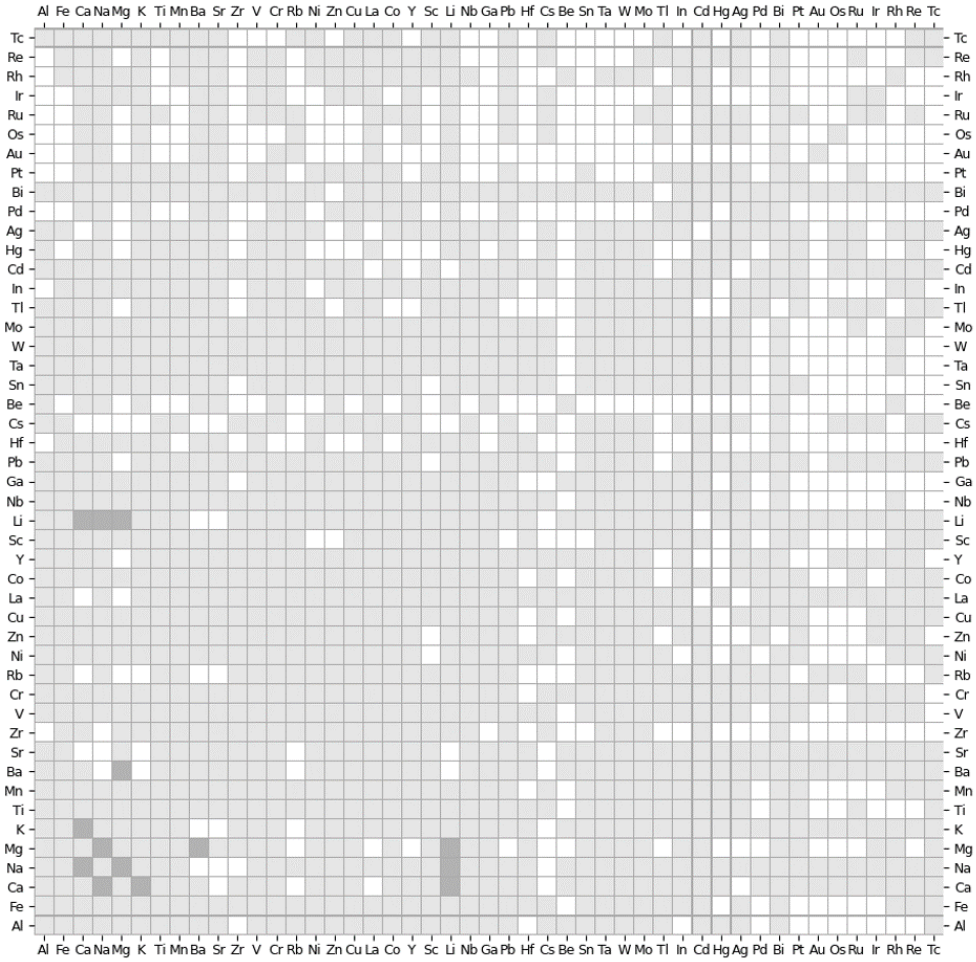

Fig. 4.1: Representation of missing ternary metal oxides. White means that the combinations of the compound are missing, gray means that at least one compound exists at the corresponding combination combinations, and black means that the whole phase diagram information exists in the database.

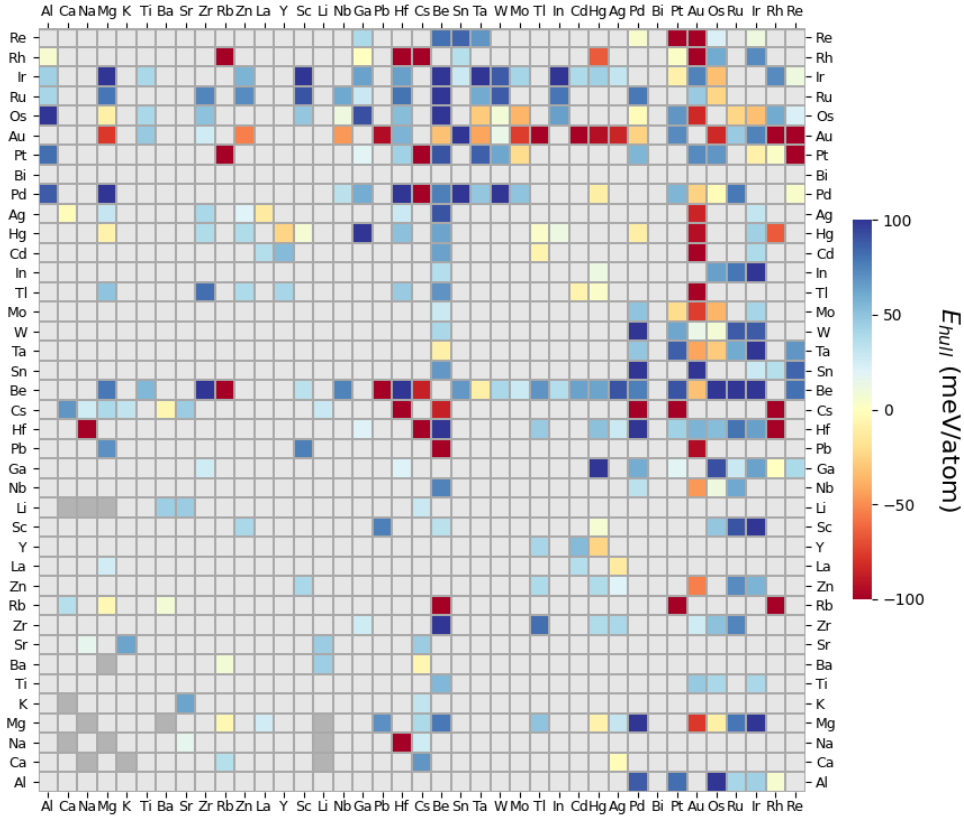

Fig. 4.2: Hull energies of the most stable structures identified by SPINNER.

4.2 Li superionic solid electrolyte

Discovering Li superconducting materials is the most essential part of developing all-solid-state batteries [119]. Until now, the best-known candidate is $Li_{10}GeP_2S_{12}$ (LGPS) because of the good Li conductivity and good mechanical stability [2]. However, this material is vulnerable to water because H_2O causes degradation of LGPS and produces toxic H_2S gas [120]. Oxide materials are much more stable than sulfides but show much poorer Li conductivity than sulfides [121]. Therefore, novel Li superionic conducting oxides with higher Li conductivity would be necessary.

Crystal structure prediction can boost identifying novel superionic conducting materials. We test SPINNER on existing Li or Na superionic conducting materials. Figure 4.3 shows the test materials. SPINNER successfully identifies all materials except for $Li_7La_3Zr_2O_{12}$ (LLZO). LLZO has 96 atoms in the primitive unit cell due to the complicated composition. We find that it takes too much time on generating structures and performing local relaxations for LLZO. So that the future work would be the fine-tuning of the algorithm to find complicated Li superionic conductors.

When trying to perform CSP on novel composition, one also faces another challenge in choosing composition because a large degree of freedom in composition exists in quaternary materials. We use design principles on composition suggested in ref. [122]: the fact that the all known promising oxide Li superionic conductor have compositions as $Li_xA_y(MO_{6/2})_m(TO_{4/2})_n$ where M and T are the metal elements that form octahedral and tetrahedral, respectively, and Li and A are the metals that incorporate into the empty space in between the corner-sharing $(MO_{6/2})_m(TO_{4/2})_n$ framework. (Note that M also can be Li.) For most oxides Li superionic conductors, m:n is typically 1:1, 1:2, 2:1, 2:3, and 3:2. For instance, $La_{3x}La_{2/3-x}TiO_3$ with NASICON structure and LLZO correspond to the case of m = 3 and m = 2. Ref. [122] also found new material, $LiTa_2PO_8$, in m = 1 and n = 2. This design principle is reasonable because if the structure consists of only tetrahedral or octahedral, it would have a too compact structure so that the space for Li conduction would be limited.

As a preliminary result, we test this design principle on Li_2ZrSiO_5. Figure 4.4a shows the atomic structure discovered by SPINNER. As expected, Zr and Si are located in octahedral and tetrahedral sites, respectively, where all polyhedrons are corner-shared. As future work, we are planning to discover oxide Li superionic conductors with new structural frameworks using SPINNER.

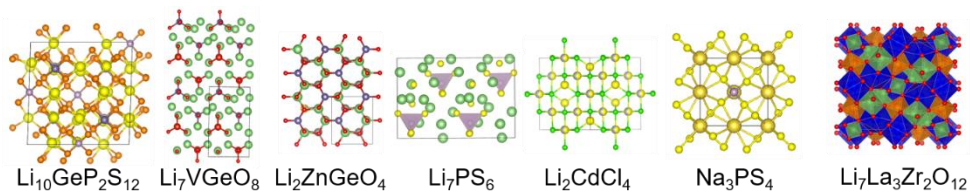

Fig. 4.3: Test materials of SPINNER for testing performance on existing Li (or Na) superionic conducting materials.

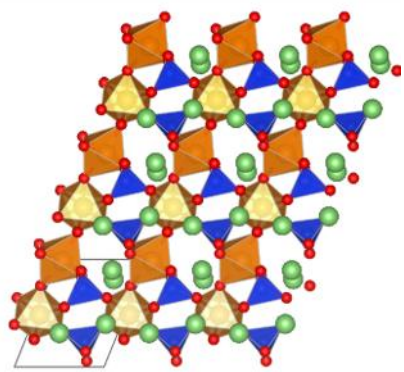

Fig. 4.4. Predicted crystal structure of Li_2ZrSiO_5.

4.3 Challenges and perspectives

In this chapter, we discuss the remaining challenges for the CSP algorithm developed in this study to be used in practical matters. The first challenge is that the algorithm is limited for predicting the crystal that consists of a large number of atoms (> 50) due to the high structure-generation and structure-relaxation times (also discussed in the previous Chapter). We think that it is hard to find general solution to this problem. Rather, one may need the case-by-case approach for each material system. For instance, in the case of Li superionic conductors, the complexity of the structure arises from the soft potential energy surface of Li. Therefore, one can generate structures without Li and put Li atoms later by Monte Carlo simulations. The second challenge is predicting crystal structures of magnetic materials. The current version of NNP is hard to directly consider the magnetic interactions. The magnetic interactions are generally local, so NNP might predict the atomic energies in the most stable spin configurations without directly encoding the spin information. However, building the training set is challenging for antiferromagnetic materials because the stable spin configurations in disordered structures are almost impossible to determine. The possible strategy is to train the initial NNP with ferromagnetic conditions and then train NNP antiferromagnetic ordered crystals in the refining stage by transfer learning. The last challenge we would like to discuss is accelerating the efficiency of CSP algorithm by integrating heuristic approaches and data-mining methods. As discussed in the Introduction section, these two methods have conflicting strengths and weaknesses: heuristic approaches are accurate but time-consuming while data-mining methods are opposite. The reason for the low accuracy of data-mining models is the deficiency of the crystal prototypes of inorganic materials in the experimental database. SPINNER can create rich prototypes in the unexplored materials domain where the data mining models are less reliable and so that can increase the accuracy of the data-mining model in diverse types of materials. Therefore, it would be possible to achieve both high accuracy and speed using a dual

model using SPINNER and a data-mining model.

Chapter 5

Conclusion

We develop the efficient and accurate crystal structure prediction (CSP) algorithm with neural network potentials (NNPs) and suggest possible applications. The NNP is trained over disordered structures constructed from molecular dynamics (MD) using density functional theory (DFT) calculations. This training procedure is tested on three ternary and one quaternary material. In all cases, the correlation between NNP and DFT energies is high enough to use NNP as a surrogate model of DFT. Then, we develop the CSP code which is named SPINNER combining iterative training scheme and quality monitoring scheme with the evolutionary searching algorithm. The program is tested over the experimental database, and successfully identifies experimental structure (or lower-energy structures) for 80% of cases. When tested over theoretical structures generated by other CSP packages, SPINNER provides the same or lower-energy structures compared to the reference structures. Finally, we apply the program for finding missing ternary metal oxides and Li superionic conductors and discuss the remaining challenges for applying SPINNER to the practical problems. By harnessing the efficiency and accuracy of NNP in CSP, this thesis will extend to the various materials discoveries.

Bibliography

[1] K. Nomura, H. Ohta, K. Ueda, T. Kamiya, M. Hirano, and H. Hosono, Science **300**, 1269 (2003).

[2] N. Kamaya, K. Homma, Y. Yamakawa, M. Hirayama, R. Kanno, M. Yonemura, T. Kamiyama, Y. Kato, S. Hama, K. Kawamoto, and A. Mitsui, Nat. Mater. **10**, 682 (2011).

[3] A. H. Slavney, T. Hu, A. M. Lindenberg, and H. I. Karunadasa, J. Am. Chem. Soc. **138**, 2138 (2016).

[4] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono, J. Am. Chem. Soc. **130**, 3296 (2008).

[5] A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov, and S. I. Shylin, Nature **525**, 73 (2015).

[6] M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, and R. J. Hemley, Phys. Rev. Lett. **122**, 027001 (2019).

[7] A. P. Drozdov, P. P. Kong, V. S. Minkov, S. P. Besedin, M. A. Kuzovnikov, S. Mozaffari, L. Balicas, F. F. Balakirev, D. E. Graf, V. B. Prakapenka, E. Greenberg, D. A. Knyazev, M. Tkacz, and M. I. Eremets, Nature **569**, 528 (2019).

[8] M. Hellenbrandt, Crystallogr. Rev. **10**, 17 (2004).

[9] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, Nat. Rev. Mater. **4**, 331 (2019).

[10] A. D. Becke, J. Chem. Phys. **140**, 18A301 (2014).

[11] Y. Kang, Y. Youn, S. Han, J. Park, and C. S. Oh, Chem. Mater. **31**, 4072 (2019).

[12] Y. Youn, M. Lee, D. Kim, J. K. Jeong, Y. Kang, and S. Han, Chem. Mater. **31**, 5475 (2019).

[13] J. Maddox, Nature **335**, 201 (1988).

[14] A. R. Oganov, *Modern Methods of Crystal Structure Prediction* (WILEY-VCH, Weinheim, 2011).

[15] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).

[16] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).

[17] W. Sun, C. J. Bartel, E. Arca, S. R. Bauers, B. Matthews, B. Orvañanos, B. R. Chen, M. F. Toney, L. T. Schelhas, W. Tumas, J. Tate, A. Zakutayev, S. Lany, A. M. Holder, and G. Ceder, Nat. Mater. **18**, 732 (2019).

[18] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier, and A. Zunger, Nat. Chem. **7**, 308 (2015).

[19] M. Nishijima, T. Ootani, Y. Kamimura, T. Sueki, S. Esaki, S. Murai, K. Fujita, K. Tanaka, K. Ohira, Y. Koyama, and I. Tanaka, Nat. Commun. **5**, 4553 (2014).

[20] C. J. Pickard and R. J. Needs, J. Phys. Condens. Matter **23**, 053201 (2011).

[21] C. J. Pickard and R. J. Needs, Phys. Rev. Lett. **97**, 045504 (2006).

[22] R. Martoňák, A. Laio, and M. Parrinello, Phys. Rev. Lett. **90**, 075503 (2003).

[23] J. Pannetier, J. Bassas-Alsina, J. Rodriquez-Carvajal, and V. Caignaert, Nature **346**, 343 (1990).
[24] S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[25] D. J. Wales and J. P. K. Doye, , J. Phys. Chem. A **101**, 5111 (1997).
[26] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Phys. Rev. B **82**, 094116 (2010).
[27] S. T. Call, D. Y. Zubarev, and A. I. Boldyrev, J. Comput. Chem. **28**, 1177 (2007).
[28] A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[29] C. W. Glass, A. R. Oganov, and N. Hansen, Comput. Phys. Commun. **175**, 713 (2006).
[30] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Comput. Phys. Commun. **183**, 2063 (2012).
[31] A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, Comput. Phys. Commun. **184**, 1172 (2013).
[32] Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, and V. Prakapenka, Nature **458**, 182 (2009).
[33] W. Zhang, A. R. Oganov, A. F. Goncharov, Q. Zhu, S. E. Boulfelfel, A. O. Lyakhov, E. Stavrou, M. Somayazulu, V. B. Prakapenka, and Z. Konopkˆová, Science **342**, 1502 (2013).
[34] H. Liu, I. I. Naumov, Z. M. Geballe, M. Somayazulu, J. S. Tse, and R. J. Hemley, Phys. Rev. B **98**, 100102(R) (2018).
[35] F. Peng, Y. Sun, C. J. Pickard, R. J. Needs, Q. Wu, and Y. Ma, Phys. Rev. Lett. **119**, 107001 (2017).
[36] D. Duan, Y. Liu, F. Tian, D. Li, X. Huang, Z. Zhao, H. Yu, B. Liu, W. Tian, and T. Cui, Sci. Rep. **4**, 6968 (2014).
[37] Y. Hinuma, T. Hatakeyama, Y. Kumagai, L. A. Burton, H. Sato, Y. Muraba, S. Iimura, H. Hiramatsu, I. Tanaka, H. Hosono, and F. Oba, Nat. Commun. **7**, 11962 (2016).
[38] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[39] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, J. Chem. Phys. **103**, 4129 (1995).
[40] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[41] V. L. Deringer, C. J. Pickard, and G. Csányi, Phys. Rev. Lett. **120**, 156001 (2018).
[42] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Phys. Rev. B **99**, 064114 (2019).
[43] Q. Tong, L. Xue, J. Lv, Y. Wang, and Y. Ma, Faraday Discuss. **211**, 31 (2018).
[44] M. Born and M. Oppenheimer, Ann. Phys. **389**, 457 (1927).
[45] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[46] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[47] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
[48] J. P. Perdew and W. Yue, Phys. Rev. B **33**, 8800(R) (1986).
[49] L. Hedin, Phys. Rev. **139**, A796 (1965).
[50] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[51] J. Sun, A. Ruzsinszky, and J. Perdew, Phys. Rev. Lett. **115**, 036402 (2015).
[52] J. H. Yang, D. A. Kitchaev, and G. Ceder, Phys. Rev. B **100**, 035132

(2019).
[53] T. Aschebrock and S. Kümmel, Phys. Rev. Res. **1**, 033082 (2019).
[54] Z. H. Yang, H. Peng, J. Sun, and J. P. Perdew, Phys. Rev. B **93**, 205205 (2016).
[55] A. V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).
[56] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Nature **589**, 59 (2021).
[57] B. Cheng, G. Mazzola, C. J. Pickard, and M. Ceriotti, Nature **585**, 217 (2020).
[58] P. Auer, H. Burgsteiner, and W. Maass, Neural Netw. **21**, 786 (2008).
[59] M. W. Mahoney and P. Drineas, Proc. Natl. Acad. Sci. U. S. A. **106**, 697 (2009).
[60] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and S. H. Sebastian, Nature **405**, 947 (2000).
[61] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[62] M. Rupp, A. Tkatchenko, K. R. Müller, and O. A. Von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[63] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K. R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[64] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[65] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, J. Chem. Phys. **148**, 241709 (2018).
[66] A. Géron, *Hands-on Machine Learning with Scikit-Learn & TensorFlow* (O'Reilly Media, Inc., Sebastopol, 2017).
[67] D. P. Kingma and J. L. Ba, *Adam: A Method for Stochastic Optimization*, ArXiv:1412.6980 (n.d.).
[68] J. C. Duchi, E. Hazan, and Y. Singer, J. Mach. Learn. Res. **12**, 2121 (2011).
[69] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization*, J. Sci. Comput. **16**, 1190 (1995).
[70] P. V. Bushlanov, V. A. Blatov, and A. R. Oganov, Comput. Phys. Commun. **236**, 1 (2019).
[71] T. Morita, J. Phys. Soc. Japan 1**2**, 753 (1957).
[72] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[73] D. W. Davies, K. T. Butler, J. M. Skelton, C. Xie, A. R. Oganov, and A. Walsh, Chem. Sci. **9**, 1022 (2018).
[74] D. C. Lonie and E. Zurek, Comput. Phys. Commun. **182**, 372 (2011).
[75] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[76] G. Bussi and A. Laio, Nat. Rev. Phys. **2**, 200 (2020).
[77] Q. Zhu, A. R. Oganov, and A. O. Lyakhov, CrystEngComm **14**, 3596 (2012).
[78] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[79] P. Shah and C. Chakravarty, Phys. Rev. Lett. **88**, 255501 (2002).
[80] S. Büchner and A. Heuer, Phys. Rev. E **60**, 6507 (1999).
[81] L. Angelani, G. Ruocco, M. Sampoli, and F. Sciortino, J. Chem. Phys. **119**, 2120 (2003).
[82] O. M. Becker, J. Chem. Physics1 **106**, 1495 (1997).

[83] A. R. Oganov and M. Valle, J. Chem. Phys. **130**, 104504 (2009).
[84] P. Avery and E. Zurek, Comput. Phys. Commun. **213**, 208 (2017).
[85] V. A. Blatov, A. P. Shevchenko, and D. M. Proserpio, Cryst. Growth Des. **14**, 3576 (2014).
[86] Z. Allahyari and A. R. Oganov, *Multi-Objective Optimization as a Tool for Material Design*, in *Handbook of Materials Modeling*, edited by Spinger (Cham, 2019).
[87] A. G. Kvashnin, C. Tantardini, H. A. Zakaryan, Y. A. Kvashnina, and A. R. Oganov, Chem. Mater. **32**, 7028 (2020).
[88] V. L. Deringer, G. Csányi, and D. M. Proserpio, ChemPhysChem **18**, 873 (2017).
[89] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[90] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[91] K. Lee, D. Yoo, W. Jeong, and S. Han, Comput. Phys. Commun. **242**, 95 (2019).
[92] D. Yoo, K. Lee, W. Jeong, D. Lee, S. Watanabe, and S. Han, Phys. Rev. Mater. **3**, 093802 (2019).
[93] S. Plimpton, J. Comp. Phys. **117**, 1 (1995).
[94] R. F. W. Herrmann, K. Tanigaki, S. Kuroshima, and H. Suematsu, Chem. Phys. Lett. **283**, 29 (1998).
[95] S. Ni, L. Chou, and J. Chang, Ceram. Int. **33**, 83 (2007).
[96] W. Weppner and R. A. Huggins, Phys. Lett. A **58**, 245 (1976).
[97] N. J. Boukharrata, J.-R. Duclère, J.-P. Laval, and P. Thomas, Acta Crystallogr. **C69**, 460 (2013).
[98] Y. Youn, M. Lee, C. Hong, D. Kim, S. Kim, J. Jung, K. Yim, and S. Han, Comput. Phys. Commun. **256**, 107450 (2020).
[99] A. O. Lyakhov, A. R. Oganov, and M. Valle, Comput. Phys. Commun. **181**, 1623 (2010).
[100] A. E. Shields, A. J. Miskowiec, K. Maheshwari, M. C. Kirkegaard, D. J. Staros, J. L. Niedziela, R. J. Kapsimalis, and B. B. Anderson, J. Phys. Chem. C **123**, 15985 (2019).
[101] I. Novikov, B. Grabowski, F. Kormann, and A. Shapeev, *Machine-Learning Interatomic Potentials Reproduce Vibrational and Magnetic Degrees of Freedom*, ArXiv:2012.127 (2020).
[102] M. Eckhoff and J. Behler, npj Comput. Mater. **7**, 170 (2021).
[103] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Mater. **1**, 011002 (2013).
[104] Y. Zhang, D. A. Kitchaev, J. Yang, T. Chen, S. T. Dacek, R. A. Sarmiento-Pérez, M. A. L. Marques, H. Peng, G. Ceder, J. P. Perdew, and J. Sun, npj Comput. Mater. **4**, 9 (2018).
[105] F. De Angelis, C. Di Valentin, S. Fantacci, A. Vittadini, and A. Selloni, Chem. Rev. **114**, 9708 (2014).
[106] Y. Wu, P. Lazic, G. Hautier, K. Persson, and G. Ceder, Energy Environ. Sci. **6**, 157 (2013).
[107] A. Narayan, A. Bhutani, S. Rubeck, J. N. Eckstein, D. P. Shoemaker, and L. K. Wagner, Phys. Rev. B **94**, 045105 (2016).
[108] N. Miao, J. Wang, Y. Gong, J. Wu, H. Niu, S. Wang, K. Li, A. R. Oganov, T. Tada, and H. Hosono, Chem. Mater. **32**, 6947 (2020).

[109] D. Luo, X. Qiao, and R. Dronskowski, Angew. Chem. Int. Ed. **60**, 486 (2021).
[110] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).
[111] X. Zhao and D. Vanderbilt, Phys. Rev. B **65**, 233106 (2002).
[112] D. Hicks, C. Toher, D. C. Ford, F. Rose, C. De Santo, O. Levy, M. J. Mehl, and S. Curtarolo, npj Comput. Mater. **7**, 30 (2021).
[113] P. Kallenbach, M. Ströbele, and H. J. Meyer, Zeitschrift Fur Anorg. Und Allg. Chemie **646**, 1281 (2020).
[114] L. Havlák, J. Fábry, M. Henriques, and M. Dusek, Acta Crystallogr. **C71**, 623 (2015).
[115] Y. Zhu, X. He, and Y. Mo, Adv. Sci. **4**, 1600517 (2017).
[116] S. Graulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, J. Appl. Crystallogr. **42**, 726 (2009).
[117] T. G. Fawcett, S. N. Kabekkodu, J. R. Blanton, and T. N. Blanton, Powder Diffr. **32**, 63 (2017).
[118] C. W. Bale, E. Bélisle, P. Chartrand, S. A. Decterov, G. Eriksson, A. E. Gheribi, K. Hack, I. H. Jung, Y. B. Kang, J. Melançon, A. D. Pelton, S. Petersen, C. Robelin, J. Sangster, P. Spencer, and M. A. Van Ende, Calphad **54**, 35 (2016).
[119] A. Manthiram, X. Yu, and S. Wang, Nat. Rev. Mater. **2**, 16103 (2017).
[120] A. Miura, N. C. Rosero-Navarro, A. Sakuda, K. Tadanaga, N. H. H. Phuc, A. Matsuda, N. Machida, A. Hayashi, and M. Tatsumisago, Nat. Rev. Chem. **3**, 189 (2019).
[121] J. C. Bachman, S. Muy, A. Grimaud, H. H. Chang, N. Pour, S. F. Lux, O. Paschos, F. Maglia, S. Lupart, P. Lamp, L. Giordano, and Y. Shao-Horn, Chem. Rev. **116**, 140 (2016).
[122] J. Kim, J. Kim, M. Avdeev, H. Yun, and S. J. Kim, J. Mater. Chem. A **6**, 22478 (2018).

초록

결정구조예측은 주어진 조성에서 가장 안정한 결정구조 상태를 찾는 방법이다. 결정구조예측 방법론을 이용한다면 원리적으로는 물질에 대한 합성 실험 이전에 합성 가능한 물질들의 라이브러리를 모두 수립할 수 있기 때문에 최근 결정구조예측 방법론은 많은 각광을 받고 있다. 하지만, 결정구조예측 방법의 한계는 알고리즘의 효율이 느리다는 것이다. 이는 결정구조예측 알고리즘이 많은 수의 제일원리계산을 동반하기 때문이다. 따라서, 제일원리계산 기반의 결정구조예측 방법은 복잡한 삼성분계 이상의 재료를 대량으로 스크리닝하는 연구들에 거의 사용되지 않고 있다. 이러한 결정구조예측 방법론의 속도를 높이기 위해서 기계학습 퍼텐셜을 제일원리계산의 대체 모델로 사용하려는 시도들이 있다. 하지만, 기계학습 퍼텐셜을 학습하기 위한 학습 데이터셋을 지정하는 것이 어려운데, 그 이유는 결정구조예측의 문제 특성상 시뮬레이션을 하기 전에 어떤 구조가 나올지 미리 알 수 없기 때문이다. 이러한 한계를 극복하기 위해 기존 연구들에서는 무작위 샘플링 방식과 실시간 학습 방식이 사용되어왔다. 하지만 기존의 이러한 방법론들은 삼성분계 이상의 시스템에 적용될 만큼 높은 정확도와 속도를 보이지는 않았다.

본 학위 논문에서는 인공신경망 퍼텐셜을 기반으로 한 결정구조예측 프로그램을 만드는 것을 목표로 하였다. 핵심 아이디어는 분자동역학 계산을 통해 만든 비정질 구조들을 인공신경망 퍼텐셜의 학습 데이터셋으로 사용하는 것이다. 이 학위논문에서는 이렇게 학습된

퍼텐셜로 계산된 에너지는 제일원리계산으로 얻어진 에너지와 높은 상관관계에 있다는 것을 밝혔다. 이는 인공 신경망 모델이 제일원리계산의 대체모델로서 사용될 수 있다는 의미이다. 이러한 인공 신경망 퍼텐셜을 기반으로 결정구조예측 프로그램인 SPINNER를 개발하였다. 프로그램은 실험 구조 데이터베이스와 이론적으로 예측된 구조들에 대하여 테스트 되었으며, 테스트 결과 개발된 방법론은 가장 안정한 결정구조를 합리적인 계산 시간 안에 찾아낼 수 있다는 것을 확인하였다. 개발된 방법론을 사용하여 진행하고 있는 삼성분계의 산화물들과 리튬 고체 전해질에 대한 탐색 연구에 대해 소개하였으며 개발된 프로그램의 한계와 발전 방향에 대하여 논하였다. 본 학위논문에서 개발된 결정구조예측 알고리즘은 우수한 미래 재료 발견으로 이어질 수 있을 것으로 기대된다.

감사의 글

먼저 박사학위 과정 동안 연구를 지도해주신 한승우 교수님께 깊은 감사의 말씀을 올립니다. 교수님께서 늘 도전적인 문제들을 제시해주시고 열정적으로 지도해주신 덕분에 학위과정 동안 많은 연구경험을 쌓으며 즐겁게 공부할 수 있었습니다. 연구 환경 및 개인적인 일에도 항상 관심을 가져주시고 힘이 되어주신 점 감사드립니다. 부족한 제 논문을 심사해 주시며 아낌없는 조언을 해주신 강기석 교수님, 이원보 교수님, 한상수 박사님, 임진오 박사님께도 다시 한 번 감사의 말씀을 올립니다.

그리고 연구실 동료 분들 및 졸업생 분들인 영호형, 계엽이형, 강훈이형, 주희누나, 용이형, 호철이형, 기재형, 규현이형, 동선이형, 동헌이형, 미소누나, 원석이형, 경풍, 상태형, 도연, 수연, (정)지수, 창호, 정민, 형민, 푸른하늘, 상민, 승우, 재훈, 지호, 민석, (김)지수에게도 감사의 말씀을 드립니다. 특히, 같이 결정구조예측 연구를 진행하며 이 논문이 완성되는데 있어서 큰 도움을 주었던 원석이형, 창호, 정민, 승우에게 감사합니다. 또한, 항상 연구에 큰 도움을 주시고 진로 및 개인적인 어려움에도 아낌없는 조언을 해주신 영호형께 다시 한번 감사의 말씀을 드립니다.

마지막으로 항상 지지해주시는 가족들에게 정말 감사합니다. 특히 항상 저에게 큰 힘이 되어주었던 사랑하는 아내 세진이에게 큰 감사의 말을 전합니다.