The Journal of Chemical Physics

RESEARCH ARTICLE | SEPTEMBER 01 2023

ChecMatE: A workflow package to automatically generate machine learning potentials and phase diagrams for semiconductor alloys

Yu-Xin Guo ; Yong-Bin Zhuang ; Jueli Shi ; Jun Cheng

J. Chem. Phys. 159, 094801 (2023)
https://doi.org/10.1063/5.0166858

AIP Publishing

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

ChecMatE: A workflow package to automatically generate machine learning potentials and phase diagrams for semiconductor alloys

Cite as: J. Chem. Phys. 159, 094801 (2023); doi: 10.1063/5.0166858
Submitted: 7 July 2023 • Accepted: 14 August 2023 •
Published Online: 1 September 2023

Yu-Xin Guo,[1] Yong-Bin Zhuang,[1,a)] Jueli Shi,[1] and Jun Cheng[1,2,b)]

AFFILIATIONS

[1] State Key Laboratory of Physical Chemistry of Solid Surfaces, iChEM, Department of Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China
[2] Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen 361005, China

[a)] E-Mail: robinzhuang@stu.xmu.edu.cn
[b)] Author to whom correspondence should be addressed: chengjun@xmu.edu.cn

ABSTRACT

Semiconductor alloy materials are highly versatile due to their adjustable properties; however, exploring their structural space is a challenging task that affects the control of their properties. Traditional methods rely on *ad hoc* design based on the understanding of known chemistry and crystallography, which have limitations in computational efficiency and search space. In this work, we present ChecMatE (Chemical Material Explorer), a software package that automatically generates machine learning potentials (MLPs) and uses global search algorithms to screen semiconductor alloy materials. Taking advantage of MLPs, ChecMatE enables a more efficient and cost-effective exploration of the structural space of materials and predicts their energy and relative stability with *ab initio* accuracy. We demonstrate the efficacy of ChecMatE through a case study of the $In_xGa_{1-x}N$ system, where it accelerates structural exploration at reduced costs. Our automatic framework offers a promising solution to the challenging task of exploring the structural space of semiconductor alloy materials.

I. INTRODUCTION

Semiconductor alloy materials, obtained by alloying two or more single semiconductor materials, have received tremendous attention in recent decades. These materials hold great promise for use in electronic and optoelectronic devices[1–5] due to their tunable properties. By adjusting their composition ratios, they can achieve the desired properties such as energy band structure, bandgap, and electron mobility.[6,7] These provide an efficient route for semiconductor devices to optimize their performance and expand the range of applications. A representative example is $In_xGa_{1-x}N$, a promising candidate for optoelectronic devices such as solar cell, light-emitting diodes (LEDs), and photoelectrodes[8–11] due to its tunable bandgap, ranging from 0.7 eV (near infrared) to 3.4 eV (ultraviolet).[12,13] It also possesses other excellent properties, such as high electron mobility and low dielectric constant, making it a promising material for photovoltaic devices.[14] Similarly, $In_xGa_{2-x}O_3$ offers an opportunity to tailor the bandgap and other material properties to meet different device requirements.[15–17]

The successful application of these semiconductor alloys in devices is grounded on in-depth understanding of their crystal structures for two main reasons. First, different crystal structures for alloys with the same composition can impact the optoelectronic properties significantly. Second, the appropriate choice of substrates based on crystal structures is the key to growing high-quality thin films.[14] Despite this, the crystal structures for most semiconductor alloys remain unclear. A solution to this problem is to construct the phase diagram of the alloy across the entire composition range for crystal structures and properties of interest.

To collect information of the phase diagram, the traditional approach involves the trial-and-error method based on experiment. However, the approach suffers from stringent synthesis procedures, various characterization processes, and is time consuming.[14] Moreover, some compositions are inaccessible, even for simple semiconductor alloys such as $In_xGa_{1-x}N$. Although GaN and InN are isostructural, the difference in the interatomic spacing between them leads to phase separation when producing high-quality $In_xGa_{1-x}N$ thin films across the entire composition range.[14,18–21] As

a result, constructing phase diagrams using the traditional method is both time-consuming and expensive.

In contrast, calculating their properties using first-principle methods is less expensive. However, a prerequisite for first-principles calculations is the accurate modeling of alloys. The main issue in this approach arises from the unknown crystal structures of the alloys. Previous studies of semiconductor alloys have primarily been based on the single crystal structures stored in databases such as ICSD[22] and Materials Project.[23] These structures are then alloyed using the special quasirandom structure (SQS) method[24–26] or by exhaustively enumerating substituted structures.[18] Consequently, these handcrafted structures cover a limited configurational space, which may omit important stable and meta-stable structures.

In recent years, predicting the unknown alloyed structures has been made possible, thanks to the development of global structure search methods such as basin-hopping (BH),[27] genetic algorithms (GAs),[28] particle swarm optimization (PSO),[29] and stochastic surface walking (SSW) methods.[30,31] These methods can explore various alloyed structures on the potential energy surface (PES), aiming to identify the global minimum, i.e., the most stable structure. Despite their utility,[19,32,33] these methods require the use of density functional theory (DFT) to accurately calculate the PES. However, the DFT calculations are computationally expensive, which impedes the application of global structure search algorithms in the exhaustive phase space search of such alloy materials, because the possible candidates for low-energy structures grow exponentially with the larger parameter space in semiconductor alloy materials. Therefore, finding an alternative method to construct the PES with less computational cost is necessary.

A promising substitute is machine learning potentials (MLPs).[34–37] Since the pioneering work of Blank *et al.*[38] in 1995, several types of MLPs and corresponding software packages have been developed, such as the Behler–Parrinello Neural Network (BPNN),[39] the Gaussian Approximation Potential (GAP),[40] the Deep Potential (DP),[41,42] and the Global Neural Network (GNN).[43,44] . Their construction lies on (i) a training dataset covering the representative structure configurations of the target PES and associated physical quantities with *ab initio* accuracy (energies, forces, and/or stresses), the latter serving as the labels for machine learning; (ii) a descriptor, converting atomic local environments to high-dimensional, symmetry-invariant vectors; (iii) a machine learning algorithm with the strong nonlinear fitting ability to create a one-to-one mapping between the descriptors and the labels.[45] Given a structure, MLPs can predict the energy, forces, and virials with *ab initio* accuracy and reduced computational costs. Because of the superior performance of MLPs, this method has been extensively applied to atomistic modeling.[35,46–49]

As a result, combining global structure search algorithms with MLPs is becoming a convenient and robust choice for structural prediction tasks in recent studies, accelerating the exploration of extensive structures for complex systems.[50–57] For instance, by combining the GAP with a PSO algorithm, Tong *et al.*[50] constructed an MLP for boron clusters and predicted the ground-state structures for B_{36} and B_{40} clusters with significantly reduced computational cost; based on the GNN and the SSW method, Ma *et al.*[52] constructed a thermodynamics phase diagram for Zn–Cr–O that reveals the presence of a small, stable composition island; very recently, Wang *et al.*[53] developed a DP model for the Cu clusters and combined it with a PSO algorithm to search for potentially stable Cu cluster structures. However, the data collection processes in the aforementioned studies result in datasets tens of thousands in size. This can be attributed to either the extensive, parallel DFT calculations carried out during the initial stage[52] or the failure to eliminate redundant structures after exploration in the active learning scheme.[53]

In this work, we aim to accelerate the exploration of the structural space of semiconductor alloys by extending the active learning scheme developed by Zhang and co-workers[58] to the Training-Exploration-Screening-Labeling Active learning (TESLA) scheme, where the screening step is added to remove redundant structures after exploration and, hence, to reduce the cost of labeling and the size of a dataset. Notice that the proposed scheme is applicable for other systems. Herein, we primarily focus on the semiconductor alloy materials.

In practice, the TESLA scheme typically involves dozens to thousands of DFT calculations and PES exploration tasks with different initial structures, which can be time-consuming and labor-intensive if done manually. To address this issue and carry out this procedure efficiently, we implemented it in the software package ChecMatE (Chemical Material Explorer) and provided reusable unit task modules (Sec. III). As a case study, we used the workflow to construct the structural phase diagram and calculate the properties for $In_xGa_{1-x}N$ alloy materials.

II. THE FRAMEWORK OF ChecMatE

As shown in Fig. 1, the framework of the ChecMatE mainly consists of three workflows: one core workflow, MLP generation, and two complementary workflows, data initialization, and structural exploration. This section will introduce them in order.

A. MLP generation

The MLP generation workflow aims to generate MLPs with *ab initio* accuracy. Since the accurate prediction of MLPs depends heavily on their training dataset, generating a high-quality training dataset is indispensible. To this end, the common practice is implementing the active learning scheme to automatically collect training datasets.[43,59–61] Here, we extend the scheme developed by Zhang and co-workers[58] to the Training-Exploration-Screening-Labeling Active learning (TESLA) scheme, which involves a series of successive iterations, and each iteration consists of four steps: training, exploration, screening and labeling.

1. Training

This step aims to generate an ensemble of MLPs using MLP training codes. Herein, we adopt the Deep Potential (DP) method, as implemented in DeePMD-kit.[42,62,63] The DP method considers the potential energy of a given structure as a sum of atomic contributions, $E = \sum E_i$, in which the E_i is determined by the local environment R_i of atom i within cutoff R_{cut} through two steps. First, the R_i is mapped, through an embedding network, to a descriptor D_i, which guarantees the permutational, translational, and rotational symmetries. The descriptor D_i is then mapped, through a fitting network, to atomic contribution E_i. The training step initializes the two networks with different random seeds for each MLP, to generate an ensemble of MLPs based on the same training dataset. These

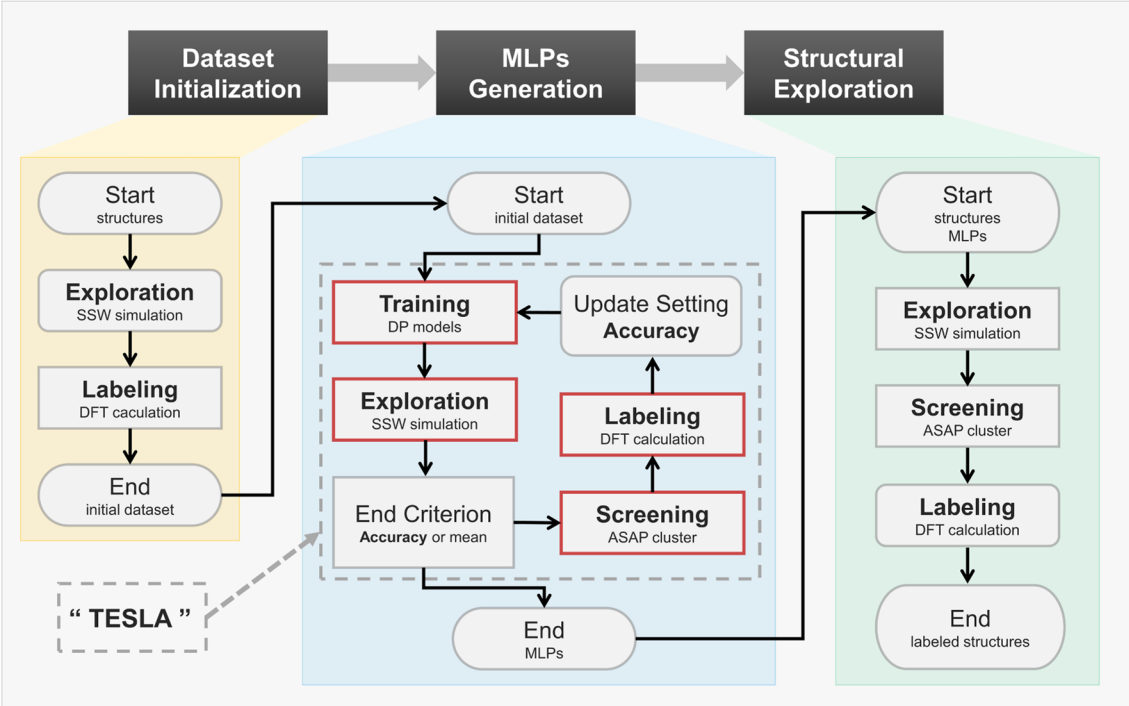

FIG. 1. Schematic diagram of the automated workflow framework, which has three processes, i.e., dataset initialization, MLP generation and structural exploration. The orange part is the dataset initialization process, where the pre-sampling trajectories are obtained by the SSW method based on DFT calculations with low precision, and a small number of structures are randomly selected to use in high-precision DFT calculations, so as to obtain the corresponding energies and atomic forces as the initial dataset. The blue part is the MLP generation process, starting with an initial dataset, which is an iterative process. In this process, an ensemble of MLPs, four by default, is constructed by the DP method. The MLPs are simultaneously used to conduct SSW simulations and calculate the model deviation for each structure on the SSW trajectories, which is the standard deviation of the maximum atomic force, i.e., $\sigma_f^{\max}$. Based on the $\sigma_f^{\max}$ and the clustering analysis implemented by ASAP software, some candidate structures are selected to be labeled by performing DFT calculations. Then, the structural energy and atomic forces of these candidate structures are added to the dataset for the next iteration. The iterations are stopped according to the accuracy or the mean of the model deviations. The green part is the structural exploration process, where an ensemble of MLPs is used along with the SSW method to explore the PES of the target systems, and low-energy structures are obtained by energy threshold and clustering screening. The low- energy structure is subjected to DFT calculations to obtain first principle results.

MLPs are used in the subsequent exploration and screening steps to explore the PES and to select outliers from the explored structures, respectively.

2. Exploration

This step aims to achieve the extension of a training dataset by efficiently exploring the PES for target systems using global search algorithms. To this end, we adopt the SSW method, as implemented in Lasphub, developed by Huang *et al.*[44] This package can greatly facilitate the PES exploration for a wide range of complex material systems by combining the SSW method with MLPs. Compared to other global search methods, such as BH, GA, and PSO, the SSW method is an unbiased, general PES search method, which can smoothly access the structures from one local minimum in the PES to another without prior knowledge. With initial structures randomly selected from the training dataset, SSW simulations are driven by the MLPs generated in the training step using the Lasphub[44] combined with the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software.[64] During the simulation processes, the physical quantities (energy, force, and/or stress) of the structures accessed are predicted by the MLPs trained in the current iteration instead of from first-principles calculations. This enables rapid and extensive exploration of the PES and reduces the associated computational cost.

3. Screening

This step aims to identify the outliers to extend the training dataset. In the exploration step, a large number of structures are generated; however not all of them are equally important for improving the quality of the MLPs. While some structures have already been well-described by the MLPs of the current iteration, adding them to the training dataset may not lead to significant improvements. Conversely, including too many similar structures in the training dataset can lead to data redundancy. Reducing data redundancy is beneficial to further decrease the computing cost and advance MLP development.[65,66] Therefore, in the screening step, we filter the explored structures based on two criteria: the model deviation of MLPs and their structural similarity.

The model deviation[58] is defined as the standard deviation of properties predicted by an ensemble of MLPs for a given structure,

which serves as the indicator, to check whether the structure is well described by the MLPs. Here, we use the maximum model deviation of the forces ($\sigma_f^{\max}$) of a structure as the indicator:

$$\sigma_f^{\max} = \max_i \sqrt{\langle \| f_i - \langle f_i \rangle \|^2 \rangle},$$

where f_i is the force component of atom i and the $\langle \cdots \rangle$ indicates the average over the ensemble of MLPs. With the user-defined upper and lower bounds, i.e., σ_f^{low} and σ_f^{high}, the explored structures are classified into accurate, candidate, and failed, according to the $\sigma_f^{\max} < \sigma_f^{low}$, $\sigma_f^{low} < \sigma_f^{\max} < \sigma_f^{high}$, and $\sigma_f^{high} < \sigma_f^{\max}$ criteria, respectively. The accurate structures are regarded as being well described by the MLPs, while the failed structures are unsuitable for labeling using first-principles calculations. Only the candidate structures are selected to extend the training dataset.

Since the search process for most global structure search algorithms, including SSW, is stochastic, many similar or identical structures are repeatedly accessed, resulting in redundant structures among the initial candidate structures. In order to remove these redundant structures, we cluster them based on the analysis of structural similarity, which is executed with the ASAP software.[67,68] The ASAP first employs the Smooth Overlap of Atomic Positions (SOAP) descriptor[69] with principal component analysis (PCA) to generate global fingerprints associated with the whole structure for the similarity measurements. Then the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm[70] is used to build clusters of similar global fingerprints. By filtering out redundant structures belonging to the same cluster, we can obtain the final candidate structures and send them to the next labeling step.

4. Labeling

This step aims to generate high-quality data using first-principles calculations of the candidate structures. For this purpose, the Vienna *Ab initio* Simulation Package (VASP)[71,72] or the CP2K software[73] can be used to calculate the energies and forces of the final candidate structures and add them to the training dataset. The updated training dataset is applied to re-train a new ensemble of MLPs in the next iteration.

The end of the TESLA iterative process is primarily determined by the accuracy, which is defined as the proportion of accurate structures to the total number of structures explored by the SSW method, as mentioned in the screening step. When the accuracy reaches a user-preset value, the ChecMatE setting parameters are updated if there is a new setting for parameters. Otherwise, the MLPs' prediction capability is considered to meet the requirements, and the iterative process ends. However, in practice, the active-learning loop will not terminate if the target accuracy is set too high to achieve or the upper and lower bounds of model deviations (σ_f^{low} and σ_f^{high}) are not well selected. The users have to stop the workflow and check the distribution of model deviations or change the settings. To realize a hand-off workflow, a backup mechanism is implemented. This mechanism uses the change in the mean of the model deviation distribution to quantitatively determine whether the iterative process needs to stop further execution.

Overall, the MLP generation workflow allows starting with a small initial dataset and ending with a high-quality dataset. It effectively reduces computational costs by mitigating the demand for an excessively large number of DFT calculations.

B. Dataset initialization

The dataset initialization is an optional and case-specific workflow that involves collecting and labeling a set of initial structures to start the MLP generation workflow. Thus, it consists of two steps: exploration and labeling, and uses the same methods and software packages as the MLP generation workflow. The difference is that no MLPs are available in this process, and the DFT calculations are used to drive the SSW simulations, which involve more computational cost compared to that of the exploration step in the MLP generation workflow. However, as only a small amount of initial structures is needed, performing the SSW simulations with a few steps is sufficient.

Additionally, as further DFT calculations in labeling are subsequently performed, the accuracy of the DFT calculations in the exploration step can be reduced to lower the computational cost. Following the exploration step, some structures are randomly sampled from the explored structures and labeled using more accurate DFT calculations. These labeled structures are then used as the initial dataset for the MLP generation workflow.

C. Structural exploration

The structural exploration workflow aims to use an ensemble of MLPs to explore the target systems' PES to find low-energy structures. This workflow consists of three steps, i.e., exploration, screening, and labeling. At this moment, all modules are reused from the MLP generation workflow; however, with slight modification. First, in the exploration step, the MLPs drive SSW simulations for a given set of initial structures to quickly explore the PES. Next, the screening step picks out the low-energy structures based on the potential energies relative to lowest values, instead of the criterion based on the model deviation, as this step is not designed for expanding the dataset. Finally, in the labeling step, the low-energy structures obtained by the screening step require DFT calculations, due to the error between the MLP predictions and DFT calculations. The results can then be used to plot the energy hull diagram and as a test dataset to further validate the accuracy of the MLPs. Compared to the MD-based exploration conducted by Zhang *et al.*,[58] the SSW method used here has higher search efficiency and reduces the required calculation cost by sifting out redundant structures.

III. SOFTWARE

A. Overview

To realize the above workflows and allocate computational resources, we implemented the ChecMatE software package using Python. Using this package, each workflow in the framework can be executed with an associated command:

```
checmate -s SETTING WORKFLOW_NAME,
```

where the argument WORKFLOW_NAME is the name of a workflow; for example, the name of the MLP generation workflow is *gen_mlps*. And the argument SETTING is the name of a user-provided parameter file in JSON or YAML format.

The ChecMatE package contains a variety of unit task modules, which can be used as components to build a specific workflow. Each unit task module, except those unrelated to computing, comprises two major parts. One part is used for processing tasks, including the reading and generation of input files and the selection of structures. The other part is a task dispatcher, used for interaction with the high-performance cluster (HPC), such as submitting and monitoring tasks. It is mainly implemented via the DPD dispatcher package,[58] generally involving the following procedures. First, based on the user's settings, the program generates the submission scripts required by the task scheduling system on the HPC. Then, all the tasks are submitted by running a command in the task directory, and a certain amount of computing resources is allocated to perform the tasks. As the tasks run, the program continuously queries the task's status until completion. Finally, the required output file is obtained, for further processing and analysis, when the task is completed. This part is the basis for the rational allocation of computing resources and can significantly improve the reusability and maintainability of task modules.

B. Unit task modules

As described in Sec. III A, unit task modules can be used as components to build a specific workflow. The automatic workflows in this paper mainly use four unit task modules: training, exploration, screening, and labeling. Except for the screening module, the remaining three unit task modules—all contain task processing and task dispatcher parts, which require the corresponding two parts of the parameters. In contrast, the screening module only needs to prepare the parameters of task processing since it only contains the task processing part. To demonstrate the usage of these modules, we use the MLP generation workflow as an example, as shown in Fig. 2, and the details for configuring the four task modules are as follows:

1. General settings

For the MLP generation workflow, the following parameters, i.e., the directory of the dataset, the order of the elemental species of structures, and the criteria for terminating the workflow, are shared for all the unit task modules. These parameters are set in the general setting part:

```
general:
  dataset: path/to/dataset
    type_map: ["In", "Ga", "N"]
    end_criterion:
      accuracy: 0.95
      mean_devi: 0.01.
```

2. Training

This module generates training tasks and the input files required by the DeePMD-kit software to obtain an ensemble of MLPs. As a result, the parameters for the training module need to contain the necessary information of the input files for the software and the configurations of computing resources required for task operation. In addition, it also needs to set the number of MLPs, i.e., the number of training tasks. Here is a simple example:

```
training:
  numb_train: 4
  dp:
    params:
      model:
        descriptor:
          type: se_e2_a
    dpdispatcher:
      machine:
        batch_type: slurm
      resources:
        number_node: 1
        group_size: 4.
```

The key `dp` is the acronym for the DeepMD-kit software, and the corresponding value of the key `params` is the parameters for the input files. The key `dpdispatcher` determines the configurations required by the DPDispatcher software, and for more details, refer to the software documentation.[74]

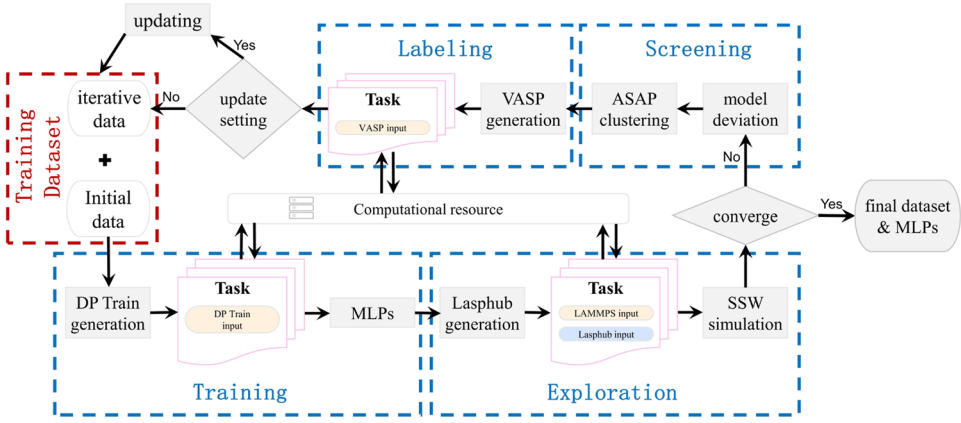

FIG. 2. Flowchart of the MLP generation workflow. The unit task modules are rounded by blue dashed lines.

3. Exploration

In this module, the exploration tasks and the input files of the Lasphub software and the LAMMPS software are prepared based on the given MLPs and the initial structures randomly selected from the dataset. Then, the SSW simulations are executed through the scheduling of computing resources. In return, the physical quantities and the model deviation of each structure explored would be calculated by the MLPs, and the results are saved in the working directory. Parameter configuration is as follows:

```
exploration:
  numb_struct_per_system: 1
  lasp:
    params:
      potential: lammps
    dpdispatcher:
      ...,
```

where the key `numb_struct_per_system` determines the number of structures selected from the systems in the dataset. The key `lasp` is the acronym of the software adopted, and here is the Lasphub software. The associated value of the key `params` is the settings required in the input file of the Lasphub software.

4. Screening

According to the model deviations, the screening module classifies all explored structures using the user-defined model deviation bounds (σ_f^{low} and σ_f^{high}). Then, structures whose model deviations fall within the bounds are extracted for clustering analysis, and candidate structures are saved in `candidates.xyz`. The relevant parameters are as follows:

```
screening:
  bounds: [0.15, 0.3]
  numb_candidate_per_traj: 25
  numb_struct_per_label: 1
  noise_percent: 100,
```

where the key `numb_candidate_per_traj` is used to set the number of candidate structures selected from every trajectory of the SSW simulations. The key `numb_struct_per_label` and `noise_percent` are the parameters related to the clustering analysis, determining the number of structures selected from the same cluster and the proportion of noise structures retained, respectively. The noise structures are those structures that cannot be clustered with their neighbors.

5. Labeling

This module generates labeling tasks and corresponding input files, which varies depending on the software. The VASP and CP2K are currently supported. In return, *ab initio* energies and atomic forces are obtained and added to the training dataset. An example of the parameters related to the VASP software is as follows:

```
labeling:
  vasp:
    params:
      incar:
        ENCUT: 750
      kpoints:
        reciprocal_density: 100
    dpdispatcher:
      ...,
```

where the key `vasp` is the name of the software adopted, and the key `params` controls the settings for DFT calculations.

Combining the parameters of the above five blocks into a single file allows one to have the parameter file required to start the MLP generation workflow. As the workflow runs, its progress is recorded in real-time in a checkpoint file based on the completion of the module. Therefore, it can be restarted from the latest progress when an error occurs or when a manual stop is intended to change parameters.

IV. EXAMPLE

Having introduced the framework and configurations for ChecMatE, we now show an example that uses the ChecMatE to construct the phase diagram of $In_xGa_{1-x}N$ across the whole concentrations.

A. Computation details

1. MLP generation

In the training step, the Deep Potential method with the smooth descriptor developed by Zhang *et al.*[41] is used to train four MLPs. The embedding network and the fitting network sizes for the MLPs are set to (25, 50, 100) and (240, 240, 240), respectively. The cutoff radius (R_{cut}) and smoothing radius required for constructing the descriptors are set to 6.0 and 0.5 Å, respectively. The training steps are 200 000, and the learning rate decays from 5×10^{-4} to 1.8×10^{-8}. During the exploration step, the SSW method in variable cell mode is used. One structure is extracted for each system in the dataset as initial structures, and the number of SSW steps is set to 100. For the screening step, the bounds of the model deviation are set to (0.1, 0.25), with 50 structures sampled from every SSW trajectory. For the labeling step, DFT calculations are performed by the VASP software, using a Perdew–Burke–Ernzerhof (PBE) functional[76,77] for exchange-correlation approximation, where a PAW pseudopotential describes the electron–particle interaction. The plane wave's kinetic energy cutoff is 750 eV, and the self-consistent-field iteration energy convergence criterion is set to 10^{-6} eV. The K-point density in the Brillouin zone is set to 100 $Å^{-3}$. Additionally, the MLP generation workflow will terminate when the accuracy of the model deviations reaches 0.98, or the change in the mean is smaller than 0.005. The rest are set by default.

2. Data initialization

In the exploration step, the number of SSW steps driven by the DFT calculations is set to two, and the plane wave's kinetic energy cutoff is 400 eV. The other configurations are the same as the those of MLP Generation.

3. Structural exploration

In the exploration step, the initial structures are customized. For the screening step, the local minima are selected from every SSW trajectory. And the energy window relative to the lowest value at each alloy concentration is set to 100 meV/atom. The other configurations are also the same as those of the MLP generation.

B. Generation of MLPs for $In_xGa_{1-x}N$

GaN and InN have the hexagonal wurtzite (W) structure and the cubic zinc-blende (ZB) structure, respectively. In contrast, the hexagonal structure is thermodynamically more stable, belonging to the $P6_3mc$ space group, so, the hexagonal structure is chosen as the configuration of initial structures. Based on this configuration, we construct a $3 \times 3 \times 2$ supercell structure of 72 atoms, which are randomly substituted, to generate initial $In_xGa_{1-x}N$ structures with different indium component concentrations. These initial structures are used to start the dataset initialization workflow, resulting in an initial dataset involving 193 structures.

Next, the MLP generation workflow begins with the initial dataset. Figure 3(a) illustrates the distribution of model deviations ($\sigma_f^{\max}$) for a total of six iterations. It can be seen that the model deviations for the $In_xGa_{1-x}N$ system are initially concentrated and have low values due to its simple structural composition. After the iter002, the distribution of model deviations tends to be unchanged. Figure 3(b) depicts that the accuracy increases and the mean value of model deviation decreases over the iterations. The workflow terminates at the iter005, since the accuracy has reached 98.6%, and the changes of the mean for model deviation have been less than 0.005. Through the MLP generation workflow, the exploration step traverses ~2.16×10^6 structures, resulting in a training dataset containing the energies and atomic forces for 2228 structures, i.e., about 0.1% of the total number of structures traversed. The widespread nature of the training dataset is shown in Fig. 3(c), which depicts the correspondence between the structure energies in the dataset and their distance-weighted Steinhart ordering parameters (OP_2), as developed by Liu and co-workers, based on the original Steinhart ordering parameter, for the sake of better distinguishing

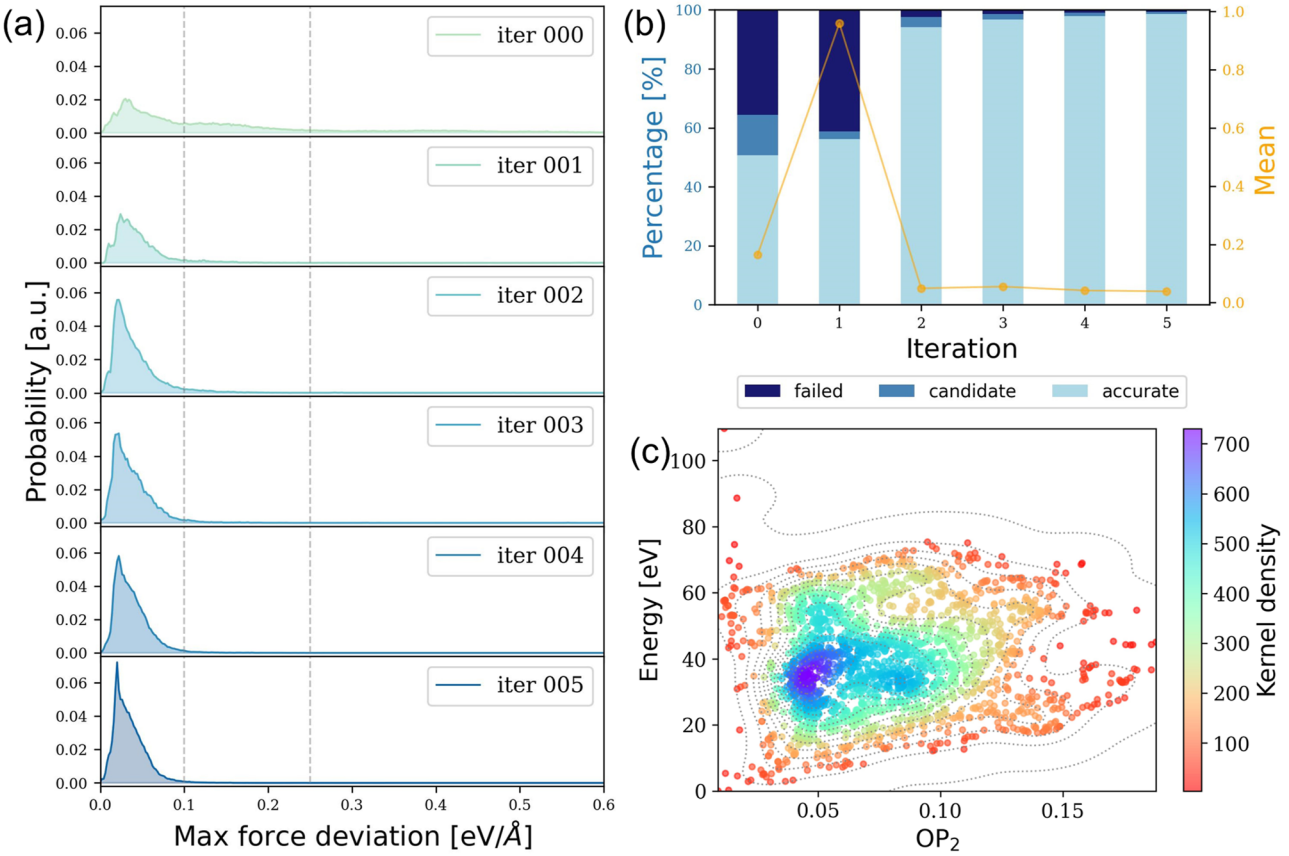

FIG. 3. The iterative processes for the $In_xGa_{1-x}N$ system. (a) Model deviation distribution diagram of all explored structures in each iteration process; the vertical dotted line represents the bounds of preset model deviation; (b) variation of accuracy: the black blue, blue gray, and blue white represent the structural proportions of failure set, candidate set, and accurate set, respectively, and the yellow dot represents the mean of the model deviation of all explored structures in the current iteration process. (c) Contour diagram of a two-dimensional PES of a training dataset. The abscissa is the distance-weighted Steinhart ordered parameter (OP_2), and the ordinate is the energy of the corresponding structure. Different colors represent the sampled structure density at the corresponding potential energy surface.

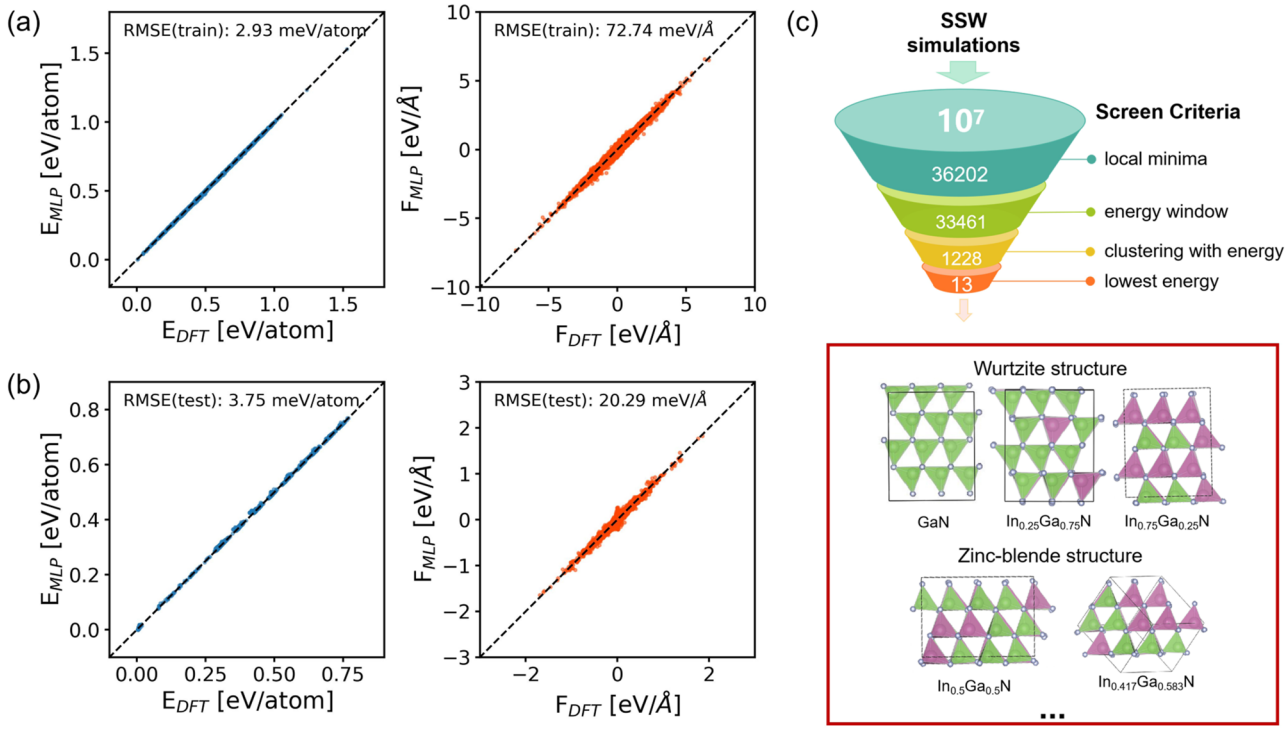

FIG. 4. Error analysis of the MLPs. The horizontal coordinate is the predicted value of the MLP model, and the vertical coordinate is the calculated value of DFT. (a) Structural energies error analysis and atomic forces error analysis for the training dataset. Each structural energy is divided by the number of atoms and subtracted from the lowest mean atomic energy; (b) structural energies error analysis and atomic forces error analysis for the test dataset. (c) Schematic diagram of the filter funnel of the structural exploration process, with each layer representing the number of structures filtered at a time. The red box shows the partial structure with the lowest formation energy at different indium concentrations, where green is gallium atom, purple is indium atom, and silver is nitrogen atom. The first layer is a hexagonal phase structure, characterized by AB-type atomic layer stacking; the second layer is a cubic phase structure, characterized by ABC-type atomic layer stacking. The structures are visualized using the VESTA software.[75]

between structures on the PES.[78,79] The OP_2 measures the short- and medium-range ordering of lattice atoms and is used to illustrate that SSW simulations explore larger areas than MD simulations.[80] Similarly, Fig. 3(c) shows that this training dataset contains the structures in a large area of the PES, not restricted to only the vicinity of a certain energy minimum of the PES. Based on this dataset, a final MLP is trained with 1 000 000 training steps. The error analysis of this MLP with respect to the training dataset is shown in Fig. 4(a). The root-mean-square errors (RMSEs) of the energies and atomic forces in the training dataset are 2.93 meV/atom and 72.74 meV/Å, respectively, indicating that the MLP is accurate for the training set.

C. Phase diagram and properties calculations

To launch the structural exploration workflow (Sec. II C), we utilize the last MLP obtained in Sec. IV B and the 293 initial structures of 13 concentrations, which evenly distributed in the range 0 to 1 for SSW simulations. For each concentration, the initial structures are extracted from the training dataset, on the condition that their potential energies per atom are less than the 100 meV/atom relative to the lowest value in the training dataset. During this workflow, SSW simulations traverse tens of millions of structures and visit 36 202 local minima of the PES. From the minima, we select 33 461 stable and meta-stable structures, whose potential energies per atom are below the 100 meV/atom relative to the global minimum for each concentration. And we then remove structurally similar structures using the cluster analysis algorithm, leaving 1228 structures. These structures serve as a test dataset to further validate the performance of the final MLP.

As shown in Fig. 4(b), the RMSEs of the structure energies and atomic forces are 3.75 meV/atom and 20.29 meV/Å, respectively, indicating that this MLP can predict these (meta-) stable structures accurately. In the final step, the most stable structure for each concentration is selected to construct the structural phase diagram, as depicted in Fig. 5(a). To be clear, we summarize the above screening steps in Fig. 4(c). The structural exploration workflow successfully obtained the most stable structure for each In concentration of $In_xGa_{1-x}N$ through the MLP accelerated global search algorithm and the step-by-step structural screening.

To better categorize the explored structures, a space group analysis is performed on these 1228 minima using ASE package,[81] and these structures are categorized into crystal systems based on the space group numbers. Due to the presence of indium in the structures, the space group analyses of the original structures are all of the low symmetry, e.g., trigonal. In order to investigate the symmetry

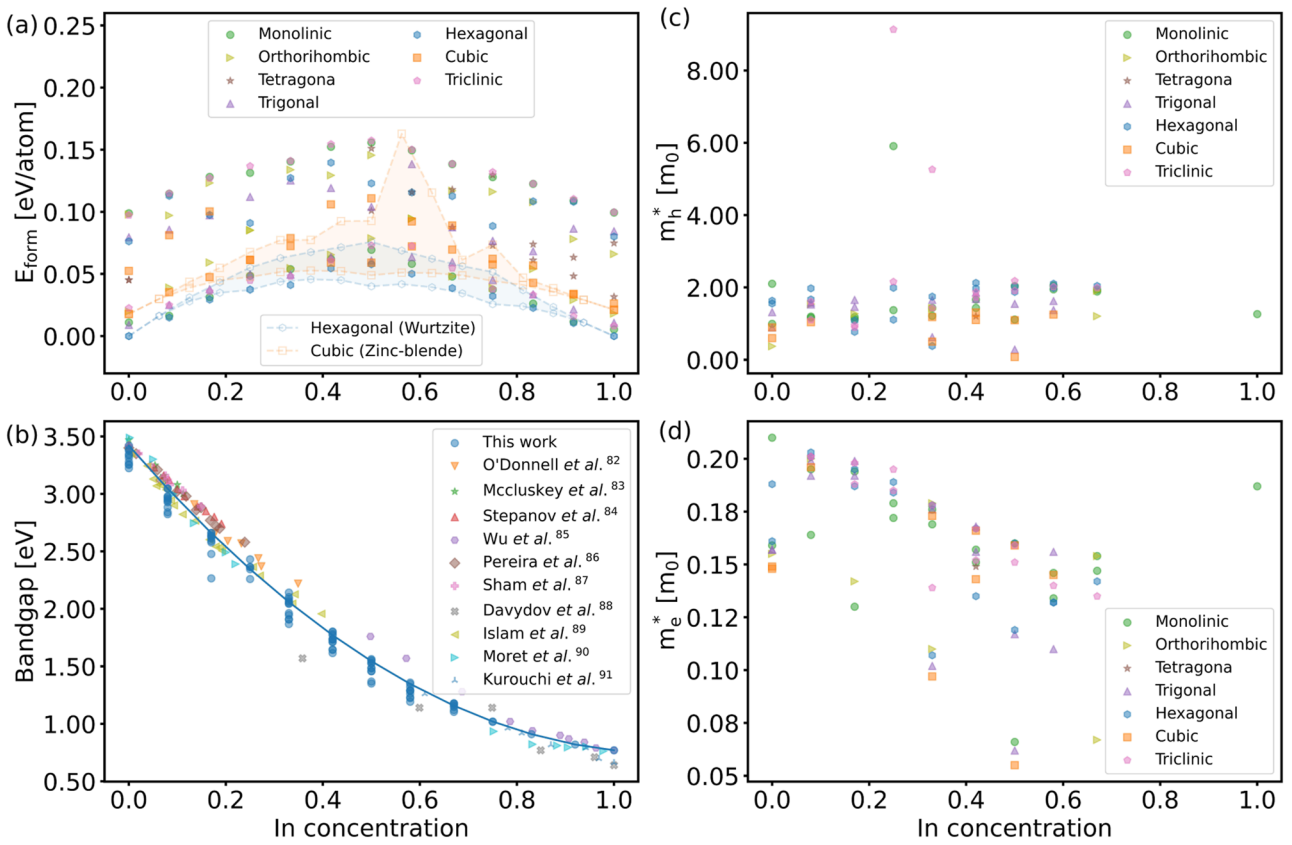

FIG. 5. (a) Formation energies predicted using MLP plotted against indium concentrations. Different markers represent various crystal systems for each concentration. Within each crystal system, only the lowest and highest formation energies are depicted. The orange and blue shaded regions represent the formation energies of all possible substituted alloys based on the experimentally observed wurtzite and zinc-blende structures. (b) DFT bandgap as a function of indium concentration and compared to the experimental data. (c) and (d) the hole and electron effective masses of the crystal systems, respectively.

of the overall structure, we ignore the element type in the structure and transform it into a structure containing a single element and then carry out a space group analysis. Meanwhile, the formation energy of each structure is calculated according to the formula, which reads

$$E_{form} = E_{In_xGa_{1-x}N} - xE_{InN} - (1-x)E_{GaN},$$

where x is the concentration of the indium component of the $In_xGa_{1-x}N$. E_{form} is the energy of formation, $E_{In_xGa_{1-x}N}$ is the energy of the $In_xGa_{1-x}N$ system, and E_{InN} and E_{GaN} are the energy of pure InN and GaN, respectively. The energy-configuration diagram of the $In_xGa_{1-x}N$ is plotted based on the formation energies and the crystal systems, together with the corresponding indium concentrations, as shown in Fig. 5(a). The lowest energy-configuration curve is a downward opening parabola, which indicates that the bulk structures of the $In_xGa_{1-x}N$ system are thermodynamically unstable and difficult to synthesize experimentally. Indeed, the low dissociation temperature of the InN component of this ternary semiconductor makes it susceptible to separation from the crystal structure.[14,20] In addition, although the workflow is based on a hexagonal phase structure to start with, other crystalline structures can still be discovered during the SSW simulations. The structure with the lowest formation energy at medium concentrations is the cubic phase structure, while at the high and low end of the alloy concentrations, it is the hexagonal phase structure, which are similar to the theoretical results of Caetano *et al.*[18]

To examine whether the global minimum of each composition is found, we calculate the formation energies of all possible substituted structures based on the experimentally observed wurtzite and zinc-blend structures, as shown in Fig. 5. The lowest formation energy of a substituted wurtzite for $In_{0.5}Ga_{0.5}N$ is found to be lower than the lowest one obtained from SSW simulations. The two structures are visualized in Fig. 6. Though both structures are wurtzite, the distribution of cations in the structure from SSW has lower symmetry compared to that in the substituted structure. Therefore, the SSW algorithm is able to find the most stable host structure; however, it does not guarantee the most favorable distribution of cations in this case. In general, we recommend users to combine the SSW simulations with the SQS method to find the most favorable distribution.

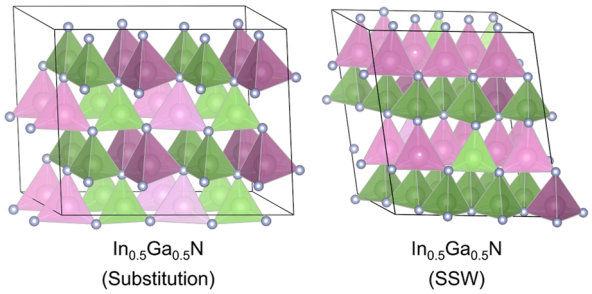

FIG. 6. The $In_{0.5}Ga_{0.5}N$ structures with the lowest formation energies from manual substitution and SSW exploration, respectively. The illustration is created using the VESTA software.[75]

Additionally, Fig. 5 shows that the lowest formation energies of cubic and hexagonal structures are very close to each other. Considering that the difference is less than the accuracy of MLP, we recalculate these formation energies using DFT and then plot them in Fig. 7. Except that the cubic structure becomes the most stable one at $x = 0.41$, the most stable structures of other compositions are impressively consistent with those predicted by the MLP.

Finally, we demonstrate that the final structures in Fig. 5(a) can be used to further extract the properties relevant to the optoelectronic application, i.e., bandgaps, and the effective masses of holes and electrons. For this purpose, we perform the first-principles calculations using the cost-effective functional, PBE, which can be replaced by more accurate functionals, such as HSE06.

Figure 5(b) shows the direct bandgap as a function of In concentrations, where we apply a linear shift[18] in the bandgaps of the alloy to make the values for the gap energy of the binary compounds GaN and InN comparable with the experimental ones (E_g^{GaN} = 3.42 eV and E_g^{InN} = 0.77 eV), respectively. Additionally, we plot other experimental results in this figure.[82–91] As shown in Fig. 5(b), we can clearly observe a downward bowing of the bandgap, and the fitted bowing parameter (b) is 2.2,

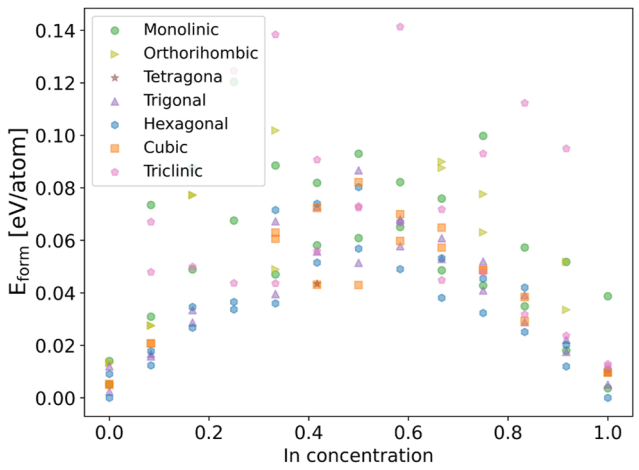

FIG. 7. DFT formation energies as a function of indium concentration. The structures are taken from Fig. 5.

which is in good agreement with the averaged experimental bowing parameter ($b = 2.5$) for relaxed thin films.[92] Figures 5(c) and 5(d) depict the smallest effective masses for holes and electrons, obtained using the parabolic band approximation, respectively, as implemented in the sumo package.[93] The effective mass, m^*, is a key parameter affecting materials' electrical properties, which is related to the carrier mobility of materials, μ, through the following formula:

$$\mu = \frac{q\tau}{m^*},$$

where the τ is the scattering time of carriers with charge q.

Currently, most studies focus on optical and structural characteristics of $In_xGa_{1-x}N$. The electrical properties of this material are relatively less investigated. For pure GaN, our calculated result for electrons is $0.16m_0$, which is in line with the experimental effective mass of $0.2m_0$.[94] Chen *et al.*[95] investigated the hole mobility of Mg-doped $In_xGa_{1-x}N$ with indium concentration ranging from 0 to 0.4. They found that carrier mobility decreased as the indium concentration was increased. The experimental results are inconsistent with our trend of the effective masses of holes, which suggests that the scattering time τ may play an important role in hole mobility. Anwar *et al.*[96] found that with increasing In concentration, the electron mobility increases, due to the decreasing effective mass. These trends are relatively consistent with our calculation results.

V. CONCLUSION

In conclusion, we have expanded upon the active learning scheme, comprised of three consecutive steps: training, exploration, and labeling. Our contribution is the inclusion of an additional step called screening, placed between exploration and labeling. This enhanced scheme, named TESLA (Training-Exploration-Screening-Labeling Active learning), has been implemented in the Python package ChecMatE (Chemical Material Explorer). It aims to facilitate the efficient collection of training datasets for machine learning potentials (MLPs) of semiconductor alloys, while minimizing data redundancy.

Furthermore, in the exploration step, we have incorporated the stochastic surface walking (SSW) method. This method generates smooth trajectories that improve the fitting of MLPs. To demonstrate the efficacy of ChecMatE, we have conducted a case study focusing on the $In_xGa_{1-x}N$ systems. By employing the TESLA scheme, we successfully converge the training dataset for these systems within six iterations, resulting in a final dataset consisting of only 2228 structures.

Using this dataset, we train an MLP capable of accurately describing the $In_xGa_{1-x}N$ systems. The MLP drives SSW simulations to explore the structures across the entire concentration range, traversing tens of millions of structures. Ultimately, we have identified 1228 meta-stable and stable structures of interest. Notably, the SSW method allows us to discover these structures without relying on prior knowledge or chemical intuition.

Additionally, we perform density functional theory calculations on the final stable and metastable structures to obtain optoelectronic properties such as band gaps and effective masses of holes and electrons. We believe that the ChecMatE package opens up new

possibilities for computational studies of semiconductor alloys, enabling accelerated material discovery and the accumulation of training data for large pretrained models.

ACKNOWLEDGMENTS

We thank Dr. Zhi-Pan Liu and Dr. Cheng Shang for addressing technical issues of the Lasphub software. Y.-X.G. acknowledges Xiamen University. Y.-B.Z. acknowledges Xiamen University and iChEM for a Ph.D. studentship (Grant Nos. 20720220007, 20720220008, 20720220009, 20720220010, and 20720220011). J.C. acknowledges the National Science Fund for Distinguished Young Scholars (Grant No. 22225302), the National Natural Science Foundation of China (Grant Nos. 21991151, 21991150, 22021001, 92161113, 91945301, and 21861132015), Laboratory of AI for Electrochemistry (AI4EC), IKKEM (Grant Nos. RD2023100101 and RD2022070501), and the Xiamen Science and Technology Plan Project (Grant No. 3502Z20203027) for financial support.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Yu-Xin Guo: Conceptualization (supporting); Formal analysis (lead); Investigation (lead); Methodology (equal); Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Yong-Bin Zhuang**: Conceptualization (lead); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (supporting). **Jueli Shi**: Conceptualization (equal); Formal analysis (supporting); Writing – original draft (supporting). **Jun Cheng**: Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Project administration (equal); Resources (lead); Supervision (equal); Writing – original draft (supporting); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. An open-source package ChecMatE is available in GitHub repository at https://github.com/chenggroup/ChecMatE.git.

REFERENCES

[1] C. S. Schnohr, Appl. Phys. Rev. **2**, 031304 (2015).
[2] T. F. Kuech, L. J. Mawst, and A. S. Brown, Annu. Rev. Chem. Biomol. Eng. **4**, 187 (2013).
[3] Y. Zhu and M. K. Hudait, Nanotechnol. Rev. **2**, 637 (2013).
[4] P. Jackson, D. Hariskos, R. Wuerz, O. Kiowski, A. Bauer, T. M. Friedlmeier, and M. Powalla, Phys. Status Solidi RRL **9**, 28 (2015).
[5] A. Chirila, P. Reinhard, F. Pianezzi, P. Bloesch, A. R. Uhl, C. Fella, L. Kranz, D. Keller, C. Gretener, H. Hagendorfer, D. Jaeger, R. Erni, S. Nishiwaki, S. Buecheler, and A. N. Tiwari, Nat. Mater. **12**, 1107 (2013).
[6] S. Adachi, *Properties of Semiconductor Alloys: Group-IV, III-V and II-VI Semiconductors* (PAMM, 2009).
[7] A. Rockett, *The Materials Science of Semiconductors* (Springer, Boston, MA, 2008).
[8] P. G. Moses and C. G. Van de Walle, Appl. Phys. Lett. **96**, 021908 (2010).
[9] P. G. Moses, M. Miao, Q. Yan, and C. G. Van de Walle, J. Chem. Phys. **134**, 084703 (2011).
[10] A. C. Meng, J. Cheng, and M. Sprik, J. Phys. Chem. B **120**, 1928 (2016).
[11] D. Wickramaratne, C. E. Dreyer, J.-X. Shen, J. L. Lyons, A. Alkauskas, and C. G. Van de Walle, Phys. Status Solidi B **257**, 1900534 (2020).
[12] F. Chen, X. Ji, and S. P. Lau, Mater. Sci. Eng., R **142**, 100578 (2020).
[13] D. Kong, Y. Zhou, J. Chai, S. Chen, L. Chen, L. Li, T. Lin, W. Wang, and G. Li, J. Mater. Chem. C **10**, 14080 (2022).
[14] R. Kour, S. Arya, S. Verma, A. Singh, P. Mahajan, and A. Khosla, ECS J. Solid State Sci. Technol. **9**, 015011 (2019).
[15] M. H. Rabbi, S. Lee, D. Sasaki, E. Kawashima, Y. Tsuruma, and J. Jang, SID Int. Symp. Dig. Tech. Pap. **53**, 16 (2022).
[16] Y. G. Kim, T. Kim, C. Avis, S.-H. Lee, and J. Jang, IEEE Trans. Electron. Devices **63**, 1078 (2016).
[17] Y. Kokubun, T. Abe, and S. Nakagomi, Phys. Status Solidi A **207**, 1741 (2010).
[18] C. Caetano, L. K. Teles, M. Marques, A. Dal Pino, and L. G. Ferreira, Phys. Rev. B **74**, 045215 (2006).
[19] D. Wines, F. Ersan, and C. Ataca, ACS Appl. Mater. Interfaces **12**, 46416 (2020).
[20] A. K. Tan, N. A. Hamzah, M. A. Ahmad, S. S. Ng, and Z. Hassan, Mater. Sci. Semicond. Process. **143**, 106545 (2022).
[21] F. K. Yam and Z. Hassan, Superlattices Microstruct. **43**, 1 (2008).
[22] M. Hellenbrandt, Crystallogr. Rev. **10**, 17 (2004).
[23] 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).
[24] H. Peelaers, D. Steiauf, J. B. Varley, A. Janotti, and C. G. Van de Walle, Phys. Rev. B **92**, 085206 (2015).
[25] S.-H. Wei, L. G. Ferreira, and A. Zunger, Phys. Rev. B **41**, 8240 (1990).
[26] M. Jaros, Rep. Prog. Phys. **48**, 1091 (1985).
[27] D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[28] S. Sivanandam and S. Deepa, *Introduction to Genetic Algorithms* (Springer, 2008).
[29] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Phys. Rev. B **82**, 094116 (2010).
[30] C. Shang and Z. P. Liu, J. Chem. Theory Comput. **9**, 1838 (2013).
[31] C. Shang, X. J. Zhang, and Z. P. Liu, Phys. Chem. Chem. Phys. **16**, 17845 (2014).
[32] D. Wines, K. Saritas, and C. Ataca, J. Chem. Phys. **155**, 194112 (2021).
[33] R. Woods-Robinson, M. K. Horton, and K. A. Persson, "A method to computationally screen for tunable properties of crystalline alloys," Patterns **4**, 100723 (2022); arXiv:2206.10715 [cond-mat].
[34] J. Behler, J. Chem. Phys. **145**, 170901 (2016).
[35] J. Behler, Chem. Rev. **121**, 10037 (2021).
[36] T. Mueller, A. Hernandez, and C. Wang, J. Chem. Phys. **152**, 050902 (2020).
[37] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J. Damewood, K. P. Greenman, and R. Gómez-Bombarelli, Acc. Mater. Res. **3**, 343 (2022).
[38] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, J. Chem. Phys. **103**, 4129 (1995).
[39] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[40] A. P. Bartok, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[41] L. Zhang, J. Han, H. Wang, W. A. Saidi, R. Car, and W. E, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems," in *Advances in Neural Information Processing Systems 31* (Curran Associates, 2019), https://proceedings.neurips.cc/paper/2018/hash/e2ad76f2326fbc6b56a45a56c59fafdb-Abstract.html.
[42] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Phys. Rev. Lett. **120**, 143001 (2018).
[43] S.-D. Huang, C. Shang, X.-J. Zhang, and Z.-P. Liu, Chem. Sci. **8**, 6327 (2017).
[44] S. Huang, C. Shang, P. Kang, X. Zhang, and Z. Liu, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **9**, e1415 (2019).

[45]J. B. Charraud, G. Geneste, M. Torrent, and J. B. Maillet, J. Chem. Phys. **156**, 204102 (2022).
[46]T. Wen, L. Zhang, H. Wang, W. E, and D. J. Srolovitz, Mater. Futures **1**, 022601 (2022).
[47]V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Chem. Rev. **121**, 10073 (2021).
[48]Y.-B. Zhuang, R.-H. Bi, and J. Cheng, J. Chem. Phys. **157**, 164701 (2022).
[49]Y.-B. Zhuang and J. Cheng, "Band alignment of metal/oxides-water interfaces using ab initio molecular dynamics," J. Electrochem. (published online) (2022).
[50]Q. Tong, L. Xue, J. Lv, Y. Wang, and Y. Ma, Faraday Discuss. **211**, 31 (2018).
[51]S.-D. Huang, C. Shang, P.-L. Kang, and Z.-P. Liu, Chem. Sci. **9**, 8644 (2018).
[52]S. Ma, S.-D. Huang, and Z.-P. Liu, Nat. Catal. **2**, 671 (2019).
[53]X. Wang, H. Wang, Q. Luo, and J. Yang, J. Chem. Phys. **157**, 074304 (2022).
[54]T. L. Jacobsen, M. S. Jørgensen, and B. Hammer, Phys. Rev. Lett. **120**, 026102 (2018).
[55]M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, and B. Hammer, J. Phys. Chem. A **122**, 1504 (2018).
[56]C. M. Andolina, P. Williamson, and W. A. Saidi, J. Chem. Phys. **152**, 154701 (2020).
[57]C. M. Andolina, J. G. Wright, N. Das, and W. A. Saidi, Phys. Rev. Mater. **5**, 083804 (2021).
[58]Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, and W. E, Comput. Phys. Commun. **253**, 107206 (2020).
[59]E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Phys. Rev. B **99**, 064114 (2019).
[60]G. Sivaraman, A. N. Krishnamoorthy, M. Baur, C. Holm, M. Stan, G. Csányi, C. Benmore, and Á. Vázquez-Mayagoitia, npj Comput. Mater. **6**, 104 (2020).
[61]C. van der Oord, M. Sachs, D. P. Kovács, C. Ortner, and G. Csányi, "Hyperactive learning (HAL) for data-driven interatomic potentials," arXiv:2210.04225 [physics.comp-ph] (2022).
[62]H. Wang, L. Zhang, J. Han, and W. E, Comput. Phys. Commun. **228**, 178 (2018).
[63]J. Zeng, D. Zhang, D. Lu, P. Mo, Z. Li, Y. Chen, M. Rynik, L. Huang, Z. Li, S. Shi, Y. Wang, H. Ye, P. Tuo, J. Yang, Y. Ding, Y. Li, D. Tisi, Q. Zeng, H. Bao, Y. Xia, J. Huang, K. Muraoka, Y. Wang, J. Chang, F. Yuan, S. L. Bore, C. Cai, Y. Lin, B. Wang, J. Xu, J.-X. Zhu, C. Luo, Y. Zhang, R. E. A. Goodall, W. Liang, A. K. Singh, S. Yao, J. Zhang, R. Wentzcovitch, J. Han, J. Liu, W. Jia, D. M. York, W. E, R. Car, L. Zhang, and H. Wang, "DeePMD-kit v2: A software package for Deep Potential models," J. Chem. Phys. **159**, 054801 (2023).
[64]A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, Comput. Phys. Commun. **271**, 108171 (2022).
[65]C. M. Andolina and W. A. Saidi, Digital Discovery **2**, 1070 (2023).
[66]P. Wisesa, C. M. Andolina, and W. A. Saidi, J. Phys. Chem. Lett. **14**, 468 (2023).
[67]B. Cheng, R. R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csányi, Acc. Chem. Res. **53**, 1981 (2020).
[68]A. Reinhardt, C. J. Pickard, and B. Cheng, Phys. Chem. Chem. Phys. **22**, 12697 (2020).
[69]A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[70]M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD'96* (AAAI Press, 1996), pp. 226–231.
[71]G. Kresse and J. Furthmuller, Phys. Rev. B **54**, 11169 (1996).
[72]G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[73]T. D. Kuhne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schutt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borstnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Muller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Gloss, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, J. Chem. Phys. **152**, 194103 (2020).
[74]https://docs.deepmodeling.com/projects/dpdispatcher/en/latest/index.html, 2020.
[75]K. Momma and F. Izumi, J. Appl. Crystallogr. **44**, 1272 (2011).
[76]J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[77]J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[78]P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[79]X. J. Zhang, C. Shang, and Z. P. Liu, Phys. Chem. Chem. Phys. **19**, 4725 (2017).
[80]P.-L. Kang, C. Shang, and Z.-P. Liu, Acc. Chem. Res. **53**, 2119 (2020).
[81]A. Hjorth Larsen, J. Jorgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dulak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiotz, O. Schutt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, J. Phys.: Condens. Matter **29**, 273002 (2017).
[82]K. O'Donnell, I. Fernandez-Torrente, P. Edwards, and R. Martin, J. Cryst. Growth **269**, 100 (2004).
[83]M. Mccluskey, C. G. Van de Walle, L. Romano, B. Krusor, and N. Johnson, J. Appl. Phys. **93**, 4340 (2003).
[84]S. Stepanov, W. Wang, B. Yavich, V. Bougrov, Y. Rebane, and Y. Shreter, MRS Internet J. Nitride Semicond. Res. **6**, e6 (2001).
[85]J. Wu, W. Walukiewicz, K. M. Yu, J. W. Ager, E. E. Haller, H. Lu, and W. J. Schaff, Appl. Phys. Lett. **80**, 4741 (2002).
[86]S. Pereira, M. R. Correia, T. Monteiro, E. Pereira, E. Alves, A. D. Sequeira, and N. Franco, Appl. Phys. Lett. **78**, 2137 (2001).
[87]W. Shan, W. Walukiewicz, E. E. Haller, B. D. Little, J. J. Song, M. D. McCluskey, N. M. Johnson, Z. C. Feng, M. Schurman, and R. A. Stall, J. Appl. Phys. **84**, 4452 (1998).
[88]V. Davydov, A. Klochikhin, V. Emtsev, D. Kurdyukov, S. Ivanov, V. Vekshin, F. Bechstedt, J. Furthmüller, J. Aderhold, J. Graul, A. Mudryi, H. Harima, A. Hashimoto, A. Yamamoto, and E. Haller, Phys. Status Solidi B **234**, 787 (2002).
[89]M. R. Islam, M. R. Kaysir, M. J. Islam, A. Hashimoto, and A. Yamamoto, J. Mater. Sci. Technol. **29**, 128 (2013).
[90]M. Moret, B. Gil, S. Ruffenach, O. Briot, C. Giesen, M. Heuken, S. Rushworth, T. Leese, and M. Succi, J. Cryst. Growth **311**, 2795 (2009).
[91]M. Kurouchi, T. Araki, H. Naoi, T. Yamaguchi, A. Suzuki, and Y. Nanishi, Phys. Status Solidi B **241**, 2843 (2004).
[92]G. Orsal, Y. El Gmili, N. Fressengeas, J. Streque, R. Djerboub, T. Moudakir, S. Sundaram, A. Ougazzaden, and J. Salvestrini, Opt. Mater. Express **4**, 1030 (2014).
[93]A. M. Ganose, A. J. Jackson, and D. O. Scanlon, J. Open Source Software **3**, 717 (2018).
[94]M. Drechsler, D. M. Hofmann, B. K. Meyer, T. Detchprohm, H. Amano, and I. Akasaki, Jpn. J. Appl. Phys. **34**, L1178 (1995).
[95]P.-C. Chen, C.-H. Chen, S.-J. Chang, Y.-K. Su, P.-C. Chang, and B.-R. Huang, Thin Solid Films **498**, 113 (2006).
[96]A. Anwar, S. Wu, and R. Webster, IEEE Trans. Electron. Devices **48**, 567 (2001).