Intricate short-range order in GeSn alloys revealed by atomistic simulations with highly accurate and efficient machine-learning potentials

Shunda Chen,[1,*] Xiaochen Jin,[1] Wanyu Zhao,[1] and Tianshu Li[1,†]

[1] *Department of Civil and Environmental Engineering, George Washington University, Washington, DC 20052, USA*

(Dated: January 8, 2024)

GeSn alloys hold promise for silicon-compatible integrated applications in electronics, photonics, and topological quantum devices. However, understanding their intricate structures using density functional theory (DFT) calculations is hindered by spatiotemporal constraints. To overcome this limitation, we develop highly accurate and efficient machine-learning interatomic potentials based on neuroevolution potential approach with active learning. The application of the developed machine-learning potential in large-scale atomistic simulations bridges the spatiotemporal gap between modeling and advanced characterizations, and facilitates the discovery of structural intricacies in GeSn alloys. Through extensive statistical sampling, we identify a new type of short-range order (SRO) which is distinguished by both its structural signature and electronic band gap from the SRO structure previously predicted. Modeling based on large simulation cell reveals the co-existence of nano SRO domains with various degrees of ordering, demonstrating a complex spatial heterogeneity of SRO structure. Our study not only reinforces the significance of fine-level structural information in alloys, but also constitutes an effective framework for exploring SRO in a broad range of complex alloys based on highly accurate and effective machine-learning potentials.

I. INTRODUCTION

In the realm of silicon-compatible integrated applications, group IV alloys have emerged as exceptionally promising materials for a wide range of applications spanning electronics, photonics, and topological quantum devices. GeSn alloys are particularly noteworthy due to their low-cost, efficiency, and tunable direct band gap, a feature crucial for mid-infrared optoelectronic applications, making them key players in the development of next-generation technologies [1–14].

Unlocking the full potential of GeSn alloys demands a comprehensive understanding of the structure-property relation, particularly in the context of atomic distribution within alloy's lattice, because the nuances in such a distribution were found capable of inducing a substantial difference in the electronic band gap [15, 16]. In this regard, although group IV alloys have been long conceived as random solid solutions, our recent theoretical studies predicted a generic and complex short-range order behavior in Si-Ge-Sn alloy systems [15–18]. Ongoing experimental characterizations also provide a strong support to the predicted structural complexity in GeSn alloys. For example, atom probe tomography (APT) measurement through an improved statistical method reveals a clear spatial fluctuation of atomic ordering in GeSn alloys [19]. Extended x-ray absorption fine structure (EXAFS) [20] shows a clear depletion of Sn-Sn first coordination number with respect to a random distribution, consistent with our prediction [15]. Polarization-dependent Raman spectroscopy probes the distinctive spectral features of Ge and Sn atoms induced by atomic ordering [21].

Despite the success in unravelling the structural complexity in GeSn alloys, a univocal interpretation of SRO requires a side-by-side comparison between modeling and characterization of SRO on the exactly same length scale. Although advanced characterization techniques such as APT [19] and energy-filtered transmission electron microscopy [22] are capable of probing SRO within a domain of ∼10 nm, this is still significantly beyond the length scale of density functional theory (DFT) based sampling approach. A paradigm shift is thus needed to bridge the existing spatiotemporal gap between modeling and characterization. This shift involves developing methodologies that transcend the spatiotemporal limitations of DFT, providing a more expansive and accurate representation of the structural intricacies inherent in GeSn alloys.

Predictive atomistic simulations could play a crucial role in this endeavor. However, their efficacy relies heavily on the quality of the interatomic potential employed. Existing classical interatomic potentials often fall short of the required accuracy due to constrained functional forms. Particularly, high-quality interatomic potentials are notedly absent for Group IV alloys, limiting the effectiveness of predictive atomistic simulations. Recently, machine-learned potentials (MLPs) have emerged as an effective framework in constructing highly accurate interatomic potentials, offering flexibility in fitting parameters and overcoming the limitations of traditional classical potentials [23–28]. In this context, the recently developed neuroevolution potential (NEP) approach [29–31] combines both excellent accuracy and outstanding computational efficiency, offering a unique advantage over other state-of-the-art methods [32–38]. In particular, optimizing the theoretical formalism and efficient implementation in the GPUMD package [39] enable NEP approach to achieve unprecedented computational speeds

* shdchen@email.gwu.edu
† tsli@email.gwu.edu

for MLPs, rivaling empirical potentials and opening avenues for accurate and efficient large-scale atomistic simulations [31, 40, 41].

In this study, we develop highly accurate and efficient NEP potentials tailored for GeSn alloys, using a comprehensive DFT dataset. The application of NEP in large-scale atomistic simulations sheds new light on the intricate behaviors of SRO in GeSn alloys that was missing in DFT-based sampling studies. In particular, the significantly enhanced spatiotemporal scale of NEP enables discovering the coexistence of two types of SROs and the resulting spatial structural heterogeneity in GeSn alloys. More importantly, the developed MLP bridges the gap between modeling and characterization, opening the door to univocally deciphering SRO and setting the example for exploring SRO in other alloy systems.

II. METHODS

A. NEP4 Neural Network Architecture for Alloy Systems

We employ an advanced NEP approach called NEP4 [41] to train the MLP for GeSn alloys comprising an extensive training dataset. NEP4 includes crucial extensions to its predecessor, NEP3 [31], specifically designed for multiple-component systems and implemented in version 3.8 of the GPUMD. In NEP4, distinct neural networks (NNs) are employed for each atom type (species), ensuring consistent regression capacity even as the number of atom types grows. Multiple loss functions are utilized to optimize different subsets of the parameters, significantly accelerating the training process using evolutionary algorithms with a large number of trainable parameters.

The previous version, NEP3 [31], is a NN potential [32] that maps a descriptor vector $\mathbf{q}^i$ (with $N_{\rm des}$ components) of a central atom i to its site energy U_i. The total energy of a system of N atoms is expressed as the sum of the site energies $U = \sum_{i=1}^{N} U^i$. The machine learning (ML) model is a fully connected feedforward NN with a single hidden layer containing $N_{\rm neu}$ neurons:

$$U^i = \sum_{\mu=1}^{N_{\rm neu}} w_\mu^{(1)} \tanh\left(\sum_{\nu=1}^{N_{\rm des}} w_{\mu\nu}^{(0)} q_\nu^i - b_\mu^{(0)}\right) - b^{(1)}, \quad (1)$$

where $\tanh(x)$ is the nonlinear activation function in the hidden layer, $\mathbf{w}^{(0)}$ is the connection weight matrix from the input layer (descriptor vector) to the hidden layer (with $N_{\rm neu}$ neurons), $\mathbf{w}^{(1)}$ is the connection weight vector from the hidden layer to the output layer, $b^{(0)}$ is the bias vector in the hidden layer, and $b^{(1)}$ is the bias in the output layer. Denoting the weight and bias parameters in the NN collectively as $\mathbf{w}$, the site energy can be expressed as:

$$U^i = \mathcal{N}\left(\mathbf{w}; \mathbf{q}^i\right). \quad (2)$$

The descriptor vector $\mathbf{q}^i$ consists of a number of radial and angular components. The radial component can be expressed as

$$q_n^i = \sum_{j\neq i} g_n(r_{ij}) \quad \text{with} \quad 0 \le n \le n_{\rm max}^{\rm R}, \quad (3)$$

where r_{ij} represents the interatomic distance between atom i and atom j, with the summation over all the neighbors of atom i within a certain cutoff distance. These radial functions can be expanded on the basis of Chebyshev polynomials [42]:

$$g_n(r_{ij}) = \sum_{k=0}^{N_{\rm bas}^{\rm R}} c_{nk}^{IJ} f_k(r_{ij}), \quad \text{with} \quad (4)$$

$$f_k(r_{ij}) = \frac{1}{2}\left[T_k\left(2\left(r_{ij}/r_{\rm c}^{\rm R} - 1\right)^2 - 1\right) + 1\right] f_{\rm c}(r_{ij}), \quad (5)$$

where $T_k(x)$ is the $k^{\rm th}$ order Chebyshev polynomial of the first kind and $f_{\rm c}(r_{ij})$ is a cutoff function, similar to Tersoff cutoff function [43, 44], defined as

$$f_{\rm c}(r_{ij}) = \begin{cases} \frac{1}{2}\left[1 + \cos\left(\pi \frac{r_{ij}}{r_{\rm c}^{\rm R}}\right)\right], & r_{ij} \le r_{\rm c}^{\rm R}; \\ 0, & r_{ij} > r_{\rm c}^{\rm R}. \end{cases} \quad (6)$$

where $r_{\rm c}^{\rm R}$ is the cutoff distance of the radial descriptor components. The expansion coefficients c_{nk}^{IJ} depend on n and k and also on the types (denoted as capitals I and J) of atoms i and j. Due to the summation over neighbors, the radial descriptor components defined above are invariant with respect to permutation of atoms of the same type. These coefficients c_{nk}^{IJ}, treated as trainable parameters [30], are crucial for efficiently distinguishing between various atom pairs contributing to the descriptor.

The 3-body angular descriptor component can be expressed as:

$$q_{nl}^i = \sum_{j\neq i}\sum_{k\neq i} g_n(r_{ij}) g_n(r_{ik}) P_l(\theta_{ijk}), \quad (7)$$

where θ_{ijk} is the angle for the triplet (ijk) with i being the central atom, and $P_l(x)$ is the Legendre polynomial of order l, with $1 \le l \le l_{\rm max}^{\rm 3b}$. Here, the radial basis functions $g_n(r_{ij})$, with $0 \le n \le n_{\rm max}^{\rm A}$, can have a different cutoff distance $r_{\rm c}^{\rm A}$ and a different basis size $N_{\rm bas}^{\rm A} + 1$. Using the addition theorem of the spherical harmonics, an equivalent form for 3-body angular descriptor component that is more efficient for numerical evaluation can be obtained:

$$q_{nl}^i = \sum_{m=-l}^{l} (-1)^m A_{nlm}^i A_{nl(-m)}^i = \sum_{m=0}^{l} (2 - \delta_{0l}) |A_{nlm}^i|^2, \quad (8)$$

where

$$A^i_{nlm} = \sum_{j \neq i} g_n(r_{ij}) Y_{lm}(\theta_{ij}, \phi_{ij}) \tag{9}$$

where $Y_{lm}(\theta_{ij}, \phi_{ij})$ are the spherical harmonics with the polar angle θ_{ij} and the azimuthal angle ϕ_{ij} for the position difference vector $\boldsymbol{r}_{ij} \equiv \boldsymbol{r}_j - \boldsymbol{r}_i$ from atom i to atom j. After obtaining the 3-body descriptor components q^i_{nl}, 4-body descriptor components $q^i_{nl_1l_2l_3}$ ($1 \le l_1 \le l_2 \le l_3 \le l^{\rm 4b}_{\rm max}$) and 5-body descriptor components $q^i_{nl_1l_2l_3l_4}$ ($1 \le l_1 \le l_2 \le l_3 \le l_4 \le l^{\rm 5b}_{\rm max}$) can be derived accordingly [31]:

$$q^i_{nl_1l_2l_3} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} \sum_{m_3=-l_3}^{l_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \times A^i_{nl_1m_1} A^i_{nl_2m_2} A^i_{nl_3m_3}, \tag{10}$$

$$q^i_{nl_1l_2l_3l_4} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} \sum_{m_3=-l_3}^{l_3} \sum_{m_4=-l_4}^{l_4} \begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} A^i_{nl_1m_1} A^i_{nl_2m_2} A^i_{nl_3m_3} A^i_{nl_4m_4}, \tag{11}$$

where $\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix}$ are Wigner $3j$ symbols [36] and

$$\begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} = \sum_{L=\max\{|l_1-l_2|,|l_3-l_4|\}}^{\min\{|l_1+l_2|,|l_3+l_4|\}} \sum_{M=-L}^{L} (-1)^M \begin{pmatrix} L & l_1 & l_2 \\ -M & m_1 & m_2 \end{pmatrix} \begin{pmatrix} L & l_3 & l_4 \\ M & m_3 & m_4 \end{pmatrix}. \tag{12}$$

The radial and angular descriptor components, namely, q^i_n, q^i_{nl}, $q^i_{nl_1l_2l_3}$, and $q^i_{nl_1l_2l_3l_4}$ are fundamentally aligned with the atom cluster expansion (ACE) formalism [36]. As all these ACE-like descriptor components involve only relative coordinates, bond lengths and bond angles, they satisfy both spatial translational and rotational invariance.

Once the radial and angular descriptor components are determined, these descriptors form the input layer of a fully connected feed-forward NN. In NEP3 approach [31], the same NN parameters $\mathbf{w}$ are applied for all the atom types (species), while the descriptor parameters $\mathbf{c}^{IJ}$ depend on the atom types (species). However, as the number of atom types increases, the regression capacity of the NN model for each atom type decreases. In NEP4 approach [41], different sets of NN parameters $\mathbf{w}^I$ are utilized for each atom type I to ensure a constant regression capacity per atom type. With this extension, the site energy of the NEP4 model can be expressed as:

$$U^i = \mathcal{N}\left(\mathbf{w}^I; \mathbf{q}^i(\{\mathbf{c}^{IJ}\})\right). \tag{13}$$

For training NEP models, we use the separable natural evolution strategy (SNES) approach [45], a powerful black-box optimization algorithm suitable for problems with many possible solutions [46]. It maintains a mean value and a variance for each trainable parameter that are updated according to the rank of a population of solutions. The rank is determined according to the loss function to be minimized. The loss function L is constructed using calculated and reference data for energies, forces, and virials and is a function of the trainable parameters. More specifically, the total loss function is defined as a weighted sum of all these individual loss functions:

$$\begin{aligned} L(\mathbf{z}) = {} & \lambda_{\rm e} \left(\frac{1}{N_{\rm str}} \sum_{n=1}^{N_{\rm str}} \left(U^{\rm NEP}(n,\mathbf{z}) - U^{\rm tar}(n)\right)^2 \right)^{1/2} \\ & + \lambda_{\rm f} \left(\frac{1}{3N} \sum_{i=1}^{N} \left(\boldsymbol{F}^{\rm NEP}_i(\mathbf{z}) - \boldsymbol{F}^{\rm tar}_i\right)^2 \right)^{1/2} \\ & + \lambda_{\rm v} \left(\frac{1}{6N_{\rm str}} \sum_{n=1}^{N_{\rm str}} \sum_{\mu\nu} \left(W^{\rm NEP}_{\mu\nu}(n,\mathbf{z}) - W^{\rm tar}_{\mu\nu}(n)\right)^2 \right)^{1/2} \\ & + \lambda_1 \frac{1}{N_{\rm par}} \sum_{n=1}^{N_{\rm par}} |z_n| \\ & + \lambda_2 \left(\frac{1}{N_{\rm par}} \sum_{n=1}^{N_{\rm par}} z_n^2 \right)^{1/2}, \end{aligned} \tag{14}$$

where vector $\mathbf{z}$ is a set of $N_{\rm par}$ total parameters, $N_{\rm str}$ is the number of structures in the training dataset (if using a full batch) or the number of structures in a mini-batch and N is the total number of atoms in these structures. $U^{\rm NEP}(n,\mathbf{z})$ and $W^{\rm NEP}_{\mu\nu}(n,\mathbf{z})$ are per-atom energy and virial tensor predicted by the NEP with parameters $\mathbf{z}$ for the $n^{\rm th}$ structure, and $\boldsymbol{F}^{\rm NEP}_i(\mathbf{z})$ is the predicted force for the $i^{\rm th}$ atom. $U^{\rm tar}(n)$, $W^{\rm tar}_{\mu\nu}(n)$ and $\boldsymbol{F}^{\rm tar}_i$ are the corresponding target values. The first three terms are the loss functions for energy, force, and virial, defined as their root-mean-square errors (RMSEs) between the current NEP predictions and the target values. The last two terms are norm-1 (ℓ_1) and norm-2 (ℓ_2) regularizations, included to avoid overfitting. The coefficients ($\lambda_{\rm e}$, $\lambda_{\rm f}$, $\lambda_{\rm v}$, λ_1, and λ_2) are their corresponding weights, which are tunable hyper-parameters.

The rank (or 'fitness') is crucial in evolutionary algorithms, determining the relative weight of a solution in the population. In the NEP3 approach [31], a single loss function is used. However, using a single loss function can lead to ambiguity in rank assignment for multiple-element systems. To address this observation, the NEP4 approach [41] introduced multiple loss functions for multiple-element systems.

$$L = L\left(\{\mathbf{w}^I\}; \{\mathbf{c}^{IJ}\}\right). \tag{15}$$

Implementing the multi-loss evolutionary algorithm in NEP4 results in faster training convergence than using a single-loss function for multiple-component systems [41].

B. NEP4 Training Hyperparameters

We use GPUMD package [31, 39] v3.8 to train the NEP4 model for GeSn alloys. The cutoff radii for radial and angular descriptor parts are 7 Å and 5 Å, respectively. For the radial descriptor components, we use 5 radial functions constructed from a linear combination of 13 basis functions. For the angular descriptor components, we use 5 radial functions constructed from a linear combination of 9 basis functions. The descriptor vector for one element thus has $5 + 5 \times 5 = 30$ components. There are 30 neurons used in the hidden layer and the NN architecture for each element can be written as 30-30-1, corresponding to 960 trainable parameters. For each pair of elements, there are $5 \times 13 + 5 \times 9 = 110$ trainable descriptor parameters. The total number of trainable parameters in the NEP4 model for GeSn alloys of 2 elements is thus $960 \times 2 + 110 \times 2^2 + 1 = 2361$, where a global bias (shifting) parameter is included. $\lambda_e = 1$, $\lambda_f = 1$, and $\lambda_v = 0.1$ are adopted. And norm-1 (ℓ_1) and norm-2 (ℓ_2) regularizations are imposed, with $\lambda_1 = \lambda_2 = 0.1$, to prevent overfitting. The training using the 90% of the full dataset (as described below) is performed with a batch size of 2000 structures for 1.25 million generations (steps), which takes about three days using four A100 GPUs. With the active-learned dataset, the training can be finished on a desktop with a single RTX4080 GPU within 1 day, or even a laptop with a single GTX1650 GPU in about one week.

C. DFT Reference Data Generation

To construct the DFT reference dataset, we employ the DFT-based Metropolis Monte Carlo (MC) method [47] to sample the configurational space of GeSn alloys. This method is chosen due to its effectiveness in achieving the necessary accuracy and accounting for local relaxation effects. The acceptance probability (P) for a new configuration j, generated from a trial move in configuration i, is given by $P = \min\{1, \exp(-(E_j - E_i)/k_B T)\}$, where E_i and E_j are the total energies of configurations i and j, respectively. Here, k_B is the Boltzmann constant, and T is the temperature (set to 300 K in this study). Each trial move involves randomly selecting and swapping a pair of atoms of different species to create a new configuration j, which then undergoes full relaxation to obtain its energy E_j.

Reference data (energy, force, and virial) for the structures are generated through DFT calculations using the Vienna *ab initio* simulation package (VASP) [48], based on the projector augmented wave method [49–51]. The local density approximation (LDA) [52] is employed for the exchange-correlation functional, known for yielding the best agreement with experimental results for geometry optimization in pure Ge and Sn [53–56]. A simulation cell containing 64 atoms, obtained by replicating a conventional diamond cubic cell containing eight atoms twice along each dimension, is chosen to ensure sufficient sampling at the DFT level. Previous investigations have demonstrated that this system size is adequate for describing the SRO structures in Si-Ge-Sn alloy systems [15, 18]. A $2 \times 2 \times 2$ Monkhorst-Pack k-points grid [57] with a plane-wave cutoff energy of 300 eV is used. The conjugate-gradient algorithm is applied for structural relaxation during each energy calculation, with convergence criteria set at 10^{-4} eV and 10^{-3} eV for electronic and ionic relaxations, respectively.

MC/DFT samplings are conducted to sample the configurational space of GeSn alloys, covering a broad range of compositions, including Sn concentrations of 3.125%, 4.6875%, 6.25%, 9.375%, 12.5%, 18.75%, 25%, 31.25%, 37.5%, 43.75%, 50%, 62.5%, 75%, 87.5%, 93.75%, 95.3125%, and 96.875%. For each Sn concentration, it involves approximately 2000 MC moves starting from different initial random configurations. Each new configuration undergoes full relaxation, where both the cell geometry and atomic positions are relaxed.

The total DFT reference dataset consists of 306 677 structures, totaling 19 627 328 atoms, for GeSn alloys, covering the full range of compositions. This dataset is randomly split, with 90% designated as the training dataset and 10% as the testing dataset. The training dataset comprises 276 009 structures, totaling 17 664 576 atoms. The testing dataset includes 30 668 structures, with a total of 1 962 752 atoms.

The broad range of composition of the initial training dataset ensures a robust NEP model capable of effectively sampling configurations in GeSn alloys. Although the DFT calculations for these structures require approximately one million CPU hours [15], it is noted that the size of the most essential data for achieving the required level of accuracy, thus the required computational cost, is significantly smaller, as explained in the active learning study below.

D. Short-Range Order Parameter

Short-range order is characterized by an atomic distribution deviating from a random solid solution within a short distance. The degree of SRO can be quantified using the Warren-Cowley SRO parameter [58], defined as

$$\alpha_{ij}^m = 1 - \frac{p_{ij}^m}{c_j}, \tag{16}$$

where c_j is the concentration of species j, and p_{ij}^m is the probability of finding atomic species j in the mth coordination shell surrounding atomic species i. The SRO parameter serves as a bulk parameter to describe the overall enhancement ($\alpha_{ij}^m < 0$) or depletion ($0 < \alpha_{ij}^m < 1$) of the i-j pair at a specific shell m with respect to a random atomic distribution. This determination involves aver-

aging the number of atomic species j surrounding each atomic species i.

E. Band Structure Calculations

For band structure calculations, we utilize the modified Becke-Johnson (mBJ) exchange potential [56], implemented in the VASP code, with the c-mBJ parameter set to 1.2. This parameter set has been demonstrated to predict accurate band gaps for Si, Ge, and α-Sn, showing good agreement with experimental data and computational efficiency superior to hybrid functionals or GW methods [53, 54, 56]. We employ the spectral weight approach [59, 60] to unfold the band structures back into the first Brillouin Zone of the diamond cubic (DC) structure. The unfolding is performed using the code *fold2bloch* [60] on 128-atom supercells, transformed from the 64-atom supercells generated from the MC/DFT trajectories. The transformation matrix is set to $[[0,1,1],[1,0,1],[1,1,0]]$.

Relativistic effects, specifically spin-orbit coupling, are included in the band structure calculation. This inclusion has been shown to be crucial for accurately reproducing the band structures of Ge and α-Sn [53, 54].

III. RESULTS

A. Training and Testing of NEP for GeSn

We first train the NEP model using 90% of the original full dataset as the training dataset. The Root Mean Square Error (RMSE) and parity plots for energy and force confirm the high accuracy of this NEP model (Fig. 1**a-d**). The training converges at about 0.2 million generations/steps Fig. 1**a-b**). On the testing dataset, the predictive accuracy reaches an excellent level, achieving RMSEs of 0.3 meV/atom for energy (Fig. 1**a** and **c**), and 15.9 meV/Å for force (Fig. 1**b** and **d**), respectively. This high level of accuracy already aligns well with the requirements for investigating the structural properties of GeSn alloys.

Given the substantial size of the original DFT dataset (306 677 structures), a pertinent question arises regarding the necessity of utilizing all the training data for MLP. To address this question, we employ an active learning technique introduced in Ref. 31 to scrutinize the descriptor space of the dataset. Specifically, we perform farthest point sampling (FPS) with a minimum Euclidean distance of 0.02 to sample the descriptor space of the original full training dataset, which contains 306 677 structures (19 627 328 atoms). FPS involves selecting data points that are farthest apart from each other, allowing us to capture the essential geometric structure of the entire dataset while avoiding redundancy. The resulting dataset from FPS sampling comprises only 137 structures, representing a three-order-of-magnitude reduction from the original training dataset. This dataset is referred as the active learning dataset, as this sampling technique aids in capturing diverse and representative samples, and can be further used for constructing an efficient and representative training dataset for MLPs under various conditions. We then visualize the descriptor space for both the original full dataset and the active learning dataset using principal component analysis (PCA) as implemented in scikit-learn [61]. The active learning dataset effectively covers the original full dataset in the two-dimensional PCA descriptor space, as depicted in Fig. 1**i**.

Utilizing this active-learned dataset, we train a new NEP model using the same hyperparameters. The predictive accuracy of this active-learned NEP model on the testing dataset shows a further improvement, achieving RMSEs of 0.2 meV/atom for energy (Fig. 1**e** and **g**) and 15.0 meV/Å for force (Fig. 1**f** and **h**), respectively. The enhanced accuracy thus highlights the significance of data efficiency of the NEP approach. Subsequently, we employ this active-learned NEP model for large-scale atomistic simulations of GeSn alloys.

Importantly, the NEP MLP also vastly enhances the computational efficiency: On a 64-atom supercell, the computational cost of MC sampling based on NEP (MC/NEP) is found reduced by approximately two orders of magnitude compared to MC/DFT. Such enhancement enables a sufficient exploration of configurational space, yielding the discovery of a new SRO structure, as indicated below. On the spatial scale, the NEP can be practically applied to explore a domain size directly comparable to advanced characterization techniques such APT, for example, $10 \times 10 \times 10$ nm^3. For this size, GeSn alloy contains $\sim$ 40000 atoms, therefore MC sampling based on DFT clearly becomes unfeasible. MC/MLP based on NEP with GPU acceleration proves to be effective: With a single desktop RTX-4080 GPU, it takes less than 20 days to complete 1 million MC steps with a 64000-atom supercell.

B. Coexistence of Two Types of SRO in GeSn Alloys

To further verify the accuracy of our developed NEP MLP, we perform MC/MLP sampling based on the developed active-learned NEP MLP for GeSn alloys and compare with the previous MC/DFT investigation [15] using the same 64-atom supercell to study $Ge_{0.75}Sn_{0.25}$ alloy. As shown in Fig. 2**a**, the energy of $Ge_{0.75}Sn_{0.25}$ is found to rapidly decrease at the initial stage for all the 8 independent MC samplings. Such decrease in energy is usually indicative of the existence of an SRO, because rearranging atomic configurations leads to optimized structures which are energetically preferred over a random al-

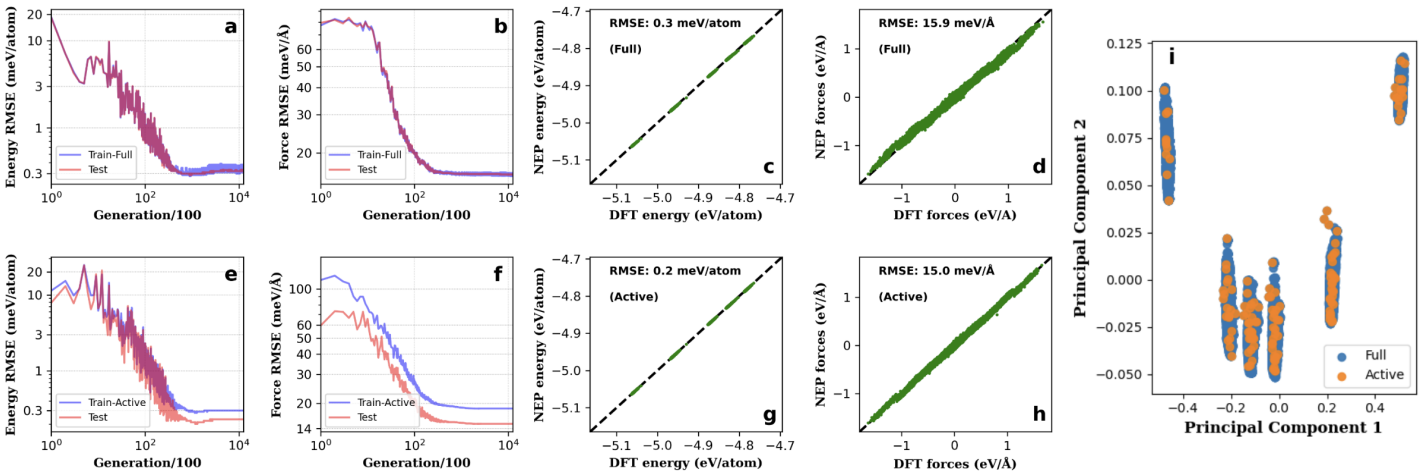

FIG. 1. Comparison of training processes using a 90% full dataset **(a-d)** and an active-learned reduced dataset **(e-h)**. The Root Mean Square Error (RMSE) for energy **(a)** and force **(b)** is illustrated during NEP training with 90% of the full dataset (276 009 structures, denoted as 'Full' in the corresponding subplots (a-d) for brevity) and the testing dataset (30 668 structures) as a function of the number of generations. Panels **(c)** and **(d)** depict the comparisons between NEP predictions and DFT reference values for energy and force, respectively, on the testing dataset (30 668 structures). In **(e-h)**, the same analyses as **(a-d)** are presented, but the model is trained with an active-learned dataset (denoted as 'Active' in the corresponding subplots **(e-h)** for brevity), which consists of a significantly reduced number of 137 structures sampled from the full dataset, as illustrated in panel **(i)** in the two-dimensional principal component descriptor space. Training with the full dataset achieves excellent accuracy, with RMSE of 0.3 meV/atom for energy and 15.9 meV/Å for force. The active-learning approach, using the same training hyperparameters, further enhances predictive accuracy on the same testing dataset (30 668 structures), achieving an RMSE of 0.2 meV/atom for energy and 15.0 meV/Å for force.

loy structure. Similar behaviors were shown in the alloys previously identified to contain SRO [15, 17, 62–64]. Importantly, MC/MLP sampling is found to faithfully reproduce the energies obtained by MC/DFT results [15] , as shown in Fig. 2**a**, thus confirming the validity of NEP MLP for studying SRO in GeSn. Employing a larger supercell, we further validate the finding by obtaining the same average 1NN Sn-Sn SRO parameters and their distribution in $Ge_{0.75}Sn_{0.25}$ alloys (see supporting information Fig. 6).

The significantly enhanced computational efficiency of NEP MLP then allows us to substantially extend the spatiotemporal scale of MC sampling, leading to discovering SRO structures that are hidden from computationally demanding, DFT-based sampling. As shown in Fig. 2**a**, the increased length of MC trajectories based on NEP MLP probes a new energy level that is lower than the energy basin identified in our previous MC/DFT study [15]. The two energy levels, denoted by high-energy basin and low-energy basin, respectively, are both well below the average energy of a random alloy, but are separated from each other by about 0.4 eV, as shown in Fig. 2**a** and Fig. 2**c**. We note that this energy difference is much greater than the typical magnitude of thermal fluctuation at room temperature. The fact that the energy hops between the two different basins may well indicate the coexistence of two distinct types of SRO, a behavior that we recently identified in Si-Ge-Sn ternary alloys [16, 18]. To verify the validity of this new energy basin discovered by NEP MLP, we further perform MC/DFT sampling starting in the structure from the low-energy basin and indeed find the sampling stays in the low-energy basin for a sizable number of MC steps (see supporting information Fig. 7).

To understand the nature of the two types of SROs in GeSn alloys, we compute the 1NN, 2NN, and 3NN Sn-Sn SRO parameters as a function of MC steps, taking MC/MLP-7 as an example. As shown in Fig. 2**a** and **b**, the 1NN, 2NN, and 3NN Sn-Sn SRO parameters show strong correlation with the total energy. In the high energy basin, the structures exhibit a strong 1NN Sn-Sn repulsion, a nearly random 2NN Sn-Sn distribution, and a mild 3NN Sn-Sn enhancement. This overall structural signature is similar to that observed in our prior study [15]. Therefore, this type of SRO is denoted as regular-SRO (R-SRO). In comparison, in the low energy basin, the structures are found to undergo further changes with respect to R-SRO, showing a complete 1NN Sn-Sn depletion, a mild 2NN Sn-Sn repulsion, and a significant 3NN Sn-Sn enhancement. As this change leads to a degree of ordering stronger than that in R-SRO, we denote it as enhanced-SRO (E-SRO).

It is noteworthy that, in the low-energy basin, a slight jump in total energy is observed, as depicted in 2**c**. Interestingly, this minor energy variation is not captured by the 1NN SRO parameter (see supporting information Fig. 8). Instead, the subtle total energy change is found well captured when including the 2NN and 3NN Sn-Sn

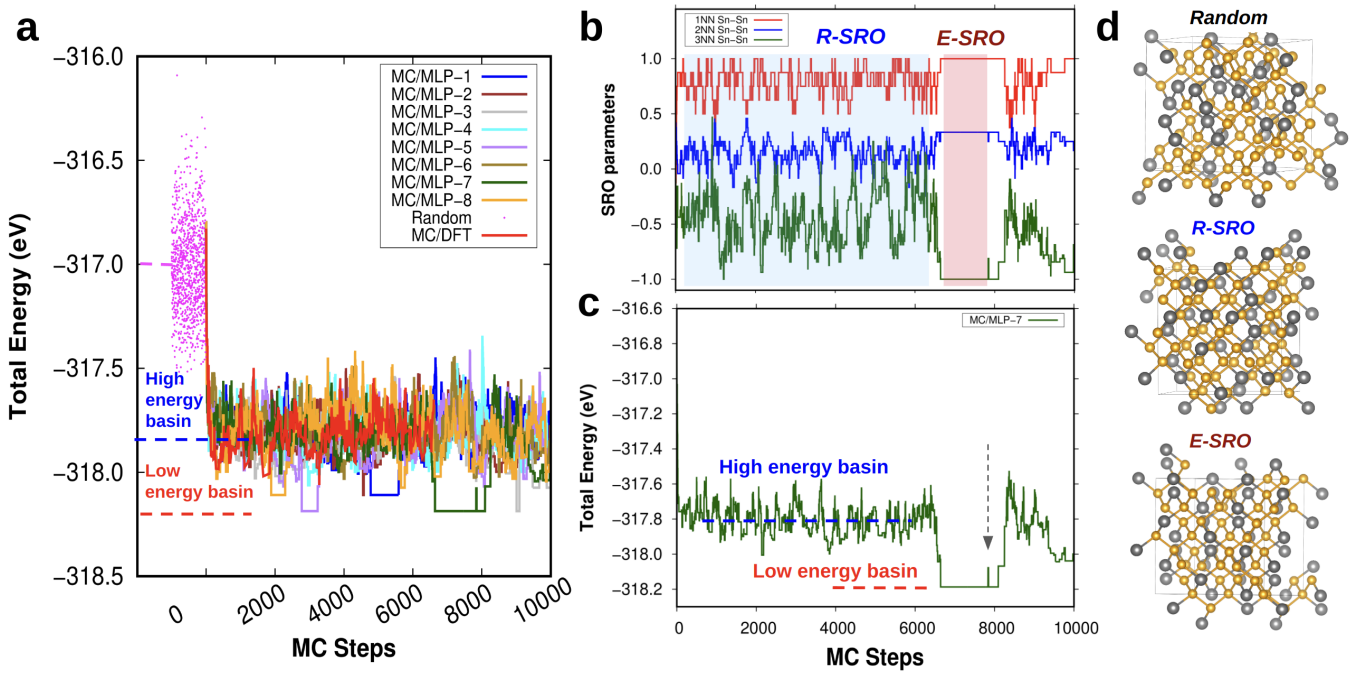

FIG. 2. Energy and structural variations arising from different types of SROs in $Ge_{0.75}Sn_{0.25}$ alloys. (**a**) The superposition of eight independent MC/MLP trajectories and one MC/DFT trajectory from our prior work [15] reveals the total energy fluctuates around the two distinct energy levels: the high-energy basin (blue dashed line) and the low-energy basin (red dashed line), separated by ∼0.4 eV. Both basins are well below the average energy (magenta dashed line) from random sampling (magenta dots). For clarity, the energy variation of MC/MLP-7 is separately shown in (**c**). Correspondingly, (**b**) illustrates the variation in the first-nearest-neighbors Sn-Sn SRO parameters (1NN, α^1_{SnSn}, red), 2NN (α^2_{SnSn}, blue), and 3NN (α^3_{SnSn}, green). Regular SRO (R-SRO) exhibits a strong 1NN Sn-Sn depletion, a mild 2NN Sn-Sn repulsion, and a moderate 3NN Sn-Sn enhancement. Enhanced SRO (E-SRO) features a complete 1NN Sn-Sn depletion, a substantial 2NN Sn-Sn repulsion, and a significant 3NN Sn-Sn enhancement. (**d**) Depicts atomic configurations of random (special quasi-random structure), R-SRO, and E-SRO structures obtained from MC/MLP-7. Ge and Sn atoms are represented by gold and gray, respectively.

SRO parameters (see Fig. 2**b**). This observation thus demonstrates the non-negligible role of 2NN and 3NN Sn-Sn interactions in shaping the intricate SRO behavior in GeSn alloys, and suggests a need to extend the conventional approach of using the average 1NN SRO parameters by considering the distributions of SRO parameters [16] or incorporating information from the further shell of nearest neighbors.

C. Impacts on Electronic Band Structures

As explored in previous studies, the presence of SRO can exert a significant influence on the electronic band structures in GeSn [15], SiSn [17], and SiGeSn alloys [16, 18]. Considering the heightened structural diversity arising from different types of SRO in GeSn alloys, a notable impact on electronic structures becomes conceivable. To investigate this possibility, we compare the calculated band structures of the $Ge_{0.75}Sn_{0.25}$ alloys exhibiting random, R-SRO, and E-SRO. As depicted in Fig. 3, significant distinctions emerge among the three structural types. The band structure of the random (SQS) alloy reveals a small direct band gap at Γ, approximately ∼0.07 eV. Introducing R-SRO increases the gap to ∼0.24 eV. Remarkably, the incorporation of E-SRO in the alloy leads to a further substantial enhancement of the direct gap at Γ to ∼0.40 eV. The detailed information of the SRO parameters and band gaps of the corresponding structures are summarized in Table I. We note this change in the band gap (0.33 eV) solely induced by SRO in $Ge_{0.75}Sn_{0.25}$, albeit not as large as that in ternary alloy $Si_{0.125}Ge_{0.625}Sn_{0.25}$, is much more significant than that (< 0.1 eV) identified in III-V semiconductor alloys exhibiting CuPt atomic ordering [65]. More importantly, the change in the band gap is sufficiently large to cross different spectral regions of infrared light, reinforcing that the nominal composition is no longer a sufficient descriptor for band gap of GeSn alloy for mid-infrared application.

TABLE I. Direct bandgaps and Sn-Sn SRO parameters (1NN: α^1_{SnSn}, 2NN: α^2_{SnSn}, and 3NN: α^3_{SnSn}) of the random (special quasi-random structure (SQS)), R-SRO, and E-SRO structures obtained from MC/MLP sampling, as depicted in Fig. 2**d**.

	Random (SQS)	R-SRO	E-SRO
Bandgap (eV)	0.07	0.24	0.40
α^1_{SnSn}	0.00	0.75	1.00
α^2_{SnSn}	0.00	0.04	0.33
α^3_{SnSn}	0.09	0.06	-1.00

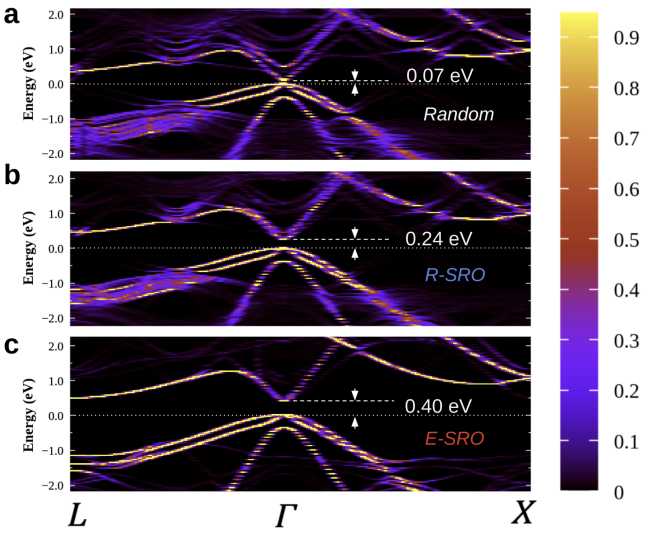

FIG. 3. Variation in the calculated electronic band structure with SRO in $Ge_{0.75}Sn_{0.25}$. (**a**) A random (SQS) distribution results in a direct band gap at Γ of 0.07 eV. (**b**) R-SRO increases the direct band gap to 0.24 eV. (**c**) E-SRO leads to a further increase of direct band gap to 0.40 eV. The corresponding Bloch spectral weight is color-coded in the bar on the right.

D. Spatial Heterogeneity of SRO Revealed by Large-scale Simulations

The significant dependence of band gap on local ordering raises a fundamental question as to which scale can GeSn alloy maintain the spatial homogeneity for a given type or a range of SRO. Such a spatial scale pertains to the size of an SRO domain within which a uniformity in atomic ordering, as reflected by, for example, the similar SRO parameters, is maintained. Relatedly, the predicted co-existence of two types of SRO is obtained using a 64-atom cell where the entire simulation cell exhibits a particular type of SRO structural character throughout a consecutive number of MC steps. It remains a question whether the co-existence manifest itself truly in real space.

To address these questions, we perform large-scale MC/MLP sampling based on NEP with GPU acceleration, for $Ge_{0.75}Sn_{0.25}$ alloys involving 64,000 atoms. The dimensions of the simulation cell amount to $11.7 \times 11.7 \times 11.7$ nm^3, already exceeding the size ($10 \times 10 \times 10$ nm^3) probed by APT [19]. As the simulation cell increases substantially, the equilibration becomes significantly longer, taking about 0.3 million MC steps for the total energy to reach equilibrium, as shown in Fig. 4 **a**. Therefore, a long MC trajectory containing $\sim$ 2.4 million steps is collected to ensure a sufficient statistical sampling.

To understand the spatial heterogeneity of SRO, we divide the simulation cell into $n \times n \times n$ domains, where n is the division number (Fig. 4 **b**). Through computing the 1NN Sn-Sn SRO parameter α^1_{SnSn} for each domain and sampling over every one hundred MC steps of the entire MC/MLP trajectory (excluding the first 300,000 steps), we construct a probability density function (PDF) for α^1_{SnSn}. As shown by Fig. 4 **c**, the ensemble-averaged 1NN Sn-Sn SRO parameter $\overline{\alpha}$ of the entire cell dimension (11.7 nm, or a division number $n = 1$) is found to be 0.78 $\pm$ 0.005, in an excellent agreement with our previous result obtained by MC/DFT sampling using 64-atom cell [15]. This not only further confirms the accuracy of the developed NEP, but also demonstrates the average SRO parameter is not subject to size effect. By dividing the simulation cell into smaller domains ($n > 1$), the corresponding PDFs for these domains are found to exhibit a clear, Gaussian distribution centered around 0.78. The width of the Gaussian peak, as reflected by the standard deviation σ, is further found to increase with n, indicating the spatial heterogeneity for the degree of SRO.

Intriguingly, when the domain size reaches $\sim$1.2 nm (division number $n = 10$), the PDF of SRO parameters displays a clear deviation from the Gaussian distribution at large SRO parameters ($\alpha^1_{\mathrm{SnSn}} > 0.85$). Remarkably, a stand-alone peak at $\alpha^1_{\mathrm{SnSn}} = 1.0$ (highlighted by a red circle) is clearly identified. Since $\alpha^1_{\mathrm{SnSn}} = 1.0$ stands for the complete depletion of Sn-Sn nearest neighbor, *i.e.*, the key structural signature of E-SRO (Fig. 2 **b**), the stand-alone peak at $\alpha^1_{\mathrm{SnSn}} = 1.0$ thus reflects the presence of the domains that only contain E-SRO. Here we note that since a division number $n = 10$ corresponds to exactly 64 atoms, namely, the cell size used in the aforementioned MC/NEP and MC/DFT simulations that predicted the co-existence of E-SRO and R-SRO (Fig. 2), the PDF analysis based on $n = 10$ thus provides the confirmation of the co-existence of E-SRO with other degrees of SRO in real space. This confirmation also suggests the co-existence of multiple degrees of SRO can be indeed reflected by simultaneous occurrence of multiple energy basins when using a small simulation cell, supporting our prior prediction of the co-existence of two SROs in Si-Ge-Sn ternary alloys [18]. The fact that the peak at $\alpha_{\mathrm{SnSn}} = 1.0$ only presents in the PDF for $n = 10$ but not for $n = 5$ indicates the domain size for E-SRO is between 1.2 nm and 2.4 nm.

Given the strong dependence of electronic band structures on the degree of SRO (Fig. 3), the spatial heterogeneity of SRO is expected to have potentially significant implications on the opto-electronic properties of

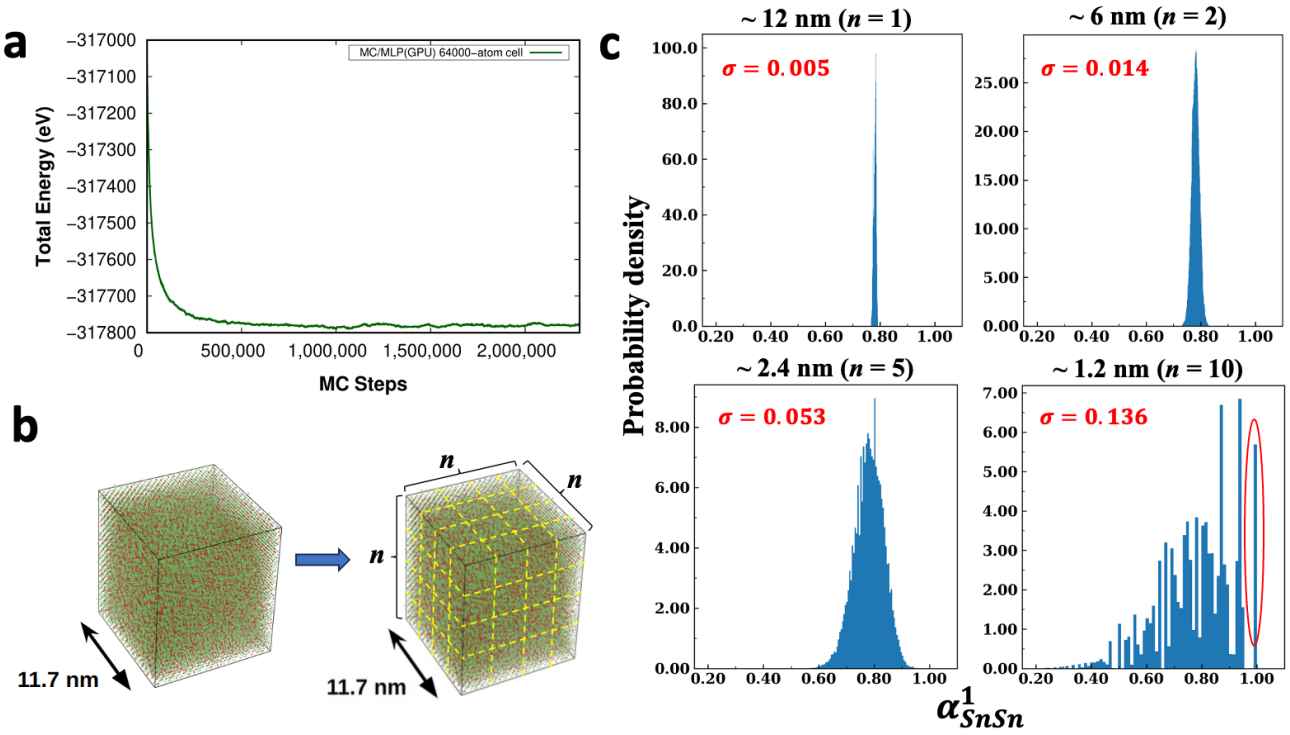

FIG. 4. **(a)** Energy variation in MC sampling using the developed NEP MLP for $Ge_{0.75}Sn_{0.25}$ alloys based on a 64,000-atom supercell with the dimension ($11.7 \times 11.7 \times 11.7$ nm^3) comparable to the volume probed by APT [19]. **(b)** The large simulation cell is evenly divided into $n \times n \times n$ domains. **(c)** The probability density function of α^1_{SnSn} for domain size of $\sim$ 12 nm (division number $n = 1$), 6 nm ($n = 2$), 2.4 nm ($n = 5$), and 1.2 nm ($n = 10$), respectively. While the average $\overline{\alpha^1_{\mathrm{SnSn}}} = 0.78$ is independent of domain size, its standard deviation σ increases with n. Particularly, for the domain size $\sim$ 1.2 nm the peak at 1.0, indicating the complete depletion of Sn-Sn bonds, is highlighted by a red circle.

Ge-Sn alloys. A likely scenario is that the SRO domains with different degrees of ordering may form natural homojunctions, due to the difference in band gaps. Since a less-ordered domain (LOD) is associated with a smaller band gap (Fig. 3), when an LOD is surrounded by more-ordered domains with greater band gaps, a nanohomojunction can form to confine the charge carriers. For $Ge_{0.75}Sn_{0.25}$ alloy, the cumulative distribution analysis shows despite the strong overall atomic ordering, there are 10% of the 1.2 nm-sized domains that have an SRO parameter less than 0.6. These LODs yield an average SRO parameter of 0.52 ± 0.08. In comparison, a neighboring domain (ND) of LOD, defined as the same-size domain that shares one face, one edge, or one corner with the LOD (schematically shown in Fig. 5**b**), is found to have an average SRO parameter of 0.81 ± 0.10, *i.e.*, 30 % further depletion of Sn-Sn bonds. Importantly, Fig. 5**b** shows the PDF of LOD and ND are virtually separated, indicating the neighboring domains of LOD are always more ordered. The analysis thus suggests there exist less-ordered, nano-sized domains that are dispersed within the overall ordered matrix, as a result of spatial heterogeneity of SRO. Since these LODs exhibit a lower band gap than their neighbors, charge carriers may be confined within these domains, resembling quantum dots, to enhance the optical response of GeSn alloy. Future studies need to be carried out to understand the optoelectronic properties of these nanohomojunctions.

IV. SUMMARY

In summary, we present the development of a highly accurate and efficient NEP MLPs to understand the SRO in GeSn alloys, leveraging a comprehensive DFT reference dataset and incorporating active learning. Our results demonstrate that active learning plays an important role. It enables a remarkable reduction of the required training data by about three orders of magnitude, making MLP training more generally accessible and feasible on standard desktop or laptop setups with a single GPU. Moreover, active learning is found to contribute to a further enhancement of predictive accuracy of MLP. This improvement may be attributed to more weighted training on representative configurations, suggesting the importance of considering the weight of representative configurations under different conditions when developing MLPs for diverse applications.

The significantly enhanced computational efficiency of NEP substantially extends the spatiotemporal scale of

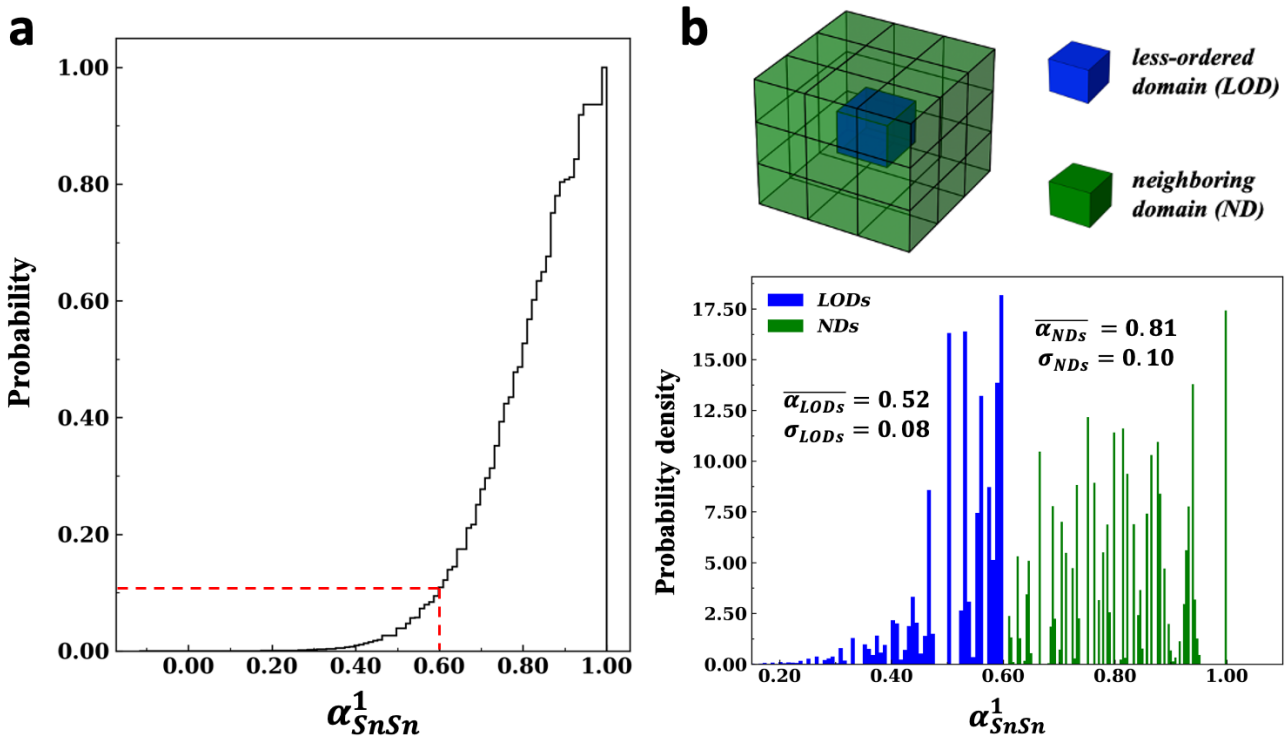

FIG. 5. **(a)** The cumulative distribution of α^1_{SnSn} for domain size $\sim$ 1.2 nm (n = 10) shows 10% of the domains yield an $\alpha^1_{\mathrm{SnSn}} < 0.6$, and these domains are called less-ordered domains (LODs). **(b)** (Top) Schematic diagram shows 26 neighboring domains (ND), colored in green, surrounding a blue-colored LOD. (Bottom) The probability density function of α^1_{SnSn} for LOD and ND are nearly separated, yielding an average $\overline{\alpha^1_{\mathrm{SnSn}}}$ of 0.52 ± 0.08 and 0.81 ± 0.10 for LOD and ND, respectively.

atomistic modeling, enabling new discovery of SRO behaviors in GeSn alloys. Applying the NEP through extensive MC sampling, we identify a new type of local ordering, E-SRO, that is featured by a complete depletion of 1NN Sn-Sn, a mild depletion of 2NN Sn-Sn, and a significant enhancement of 3NN Sn-Sn, thus exhibiting a greater degree of SRO than the previously-predicted R-SRO [15]. It can be expected that the E-SRO observed in $Ge_{0.75}Sn_{0.25}$ may be different from the E-SRO observed in $Si_{0.125}Ge_{0.625}Sn_{0.25}$ [18], in terms of 2NN and 3NN Sn-Sn. Our recent revisit of the E-SRO in $Si_{0.125}Ge_{0.625}Sn_{0.25}$ showed a complete 1NN Sn-Sn depletion, a mild 2NN Sn-Sn enhancement and a complete 3NN Sn-Sn depletion [66]. The addition of Si introduces complexity to the SRO behavior in Si-Ge-Sn alloys, which deserves further investigation using the NEP MLP framework and active learning strategy demonstrated in this work.

The structural variations induced by SRO are also found to exert significant influence on the band structures of $Ge_{0.75}Sn_{0.25}$ alloys, yielding a change in the band gap from 0.07 eV for random alloy to 0.4 eV for E-SRO. With the stability of E-SRO further confirmed by DFT calculations, both types of SRO are predicted to co-exist in GeSn alloy. Such co-existence is subsequently confirmed using a large-scale NEP modeling based on a simulation volume of $11.7 \times 11.7 \times 11.7$ nm^3, matching the effective size characterized by atom probe tomography [19]. The large-scale modeling clearly unravels the structural heterogeneity of GeSn alloy, demonstrating the presence of nano-sized SRO domains with various degrees of local ordering. In particular, this structural heterogeneity is further found to yield nanohomojunctions composed of SRO domains with different band gaps which may thus have potential interesting optoelectronic properties.

While this study is focused on GeSn alloy, several implications are broadly applicable to many other alloy systems. Firstly, the framework of NEP4 MLP can be an effective approach for exploring SRO. Although empirical interatomic potentials may be available for limited alloy systems, there is a general lack of accurate description of interatomic interactions capable of differentiating SRO for the vast majority of alloys. In particular, the subtlety of SRO often imposes a stringent requirement, both for the accuracy of potential, *e.g.*, 1 $\sim$ 10 meV/atom, and for the ability to properly account for the variation in local environment induced by SRO, *e.g.*, local distortion. The combined need for both speed and accuracy makes NEP a favorable choice for exploring SRO.

Secondly, the active learning strategy employed and demonstrated in this work for GeSn alloys can be extended to other alloys, especially more complicated ones

like high-entropy alloys. This extension can effectively reduce the size of the training dataset, lower training costs, and improve the accuracy of MLPs. It also indicates the importance of high-quality and representative training datasets for developing top-notch MLPs.

Thirdly, the identified intricacies of SRO reinforces the significance of fine-level structural information in alloys. As SRO reflects the deviation from the random distribution, the precise description of SRO should therefore mirror the nature of a statistical distribution, which should encompass at least two different senses: One refers to the frequency for a particular type of local structure to be adopted, namely, local atomic SRO distribution, as examined in our recent study [16]; The other pertains to the spatial variation of a real-space domain where a specific distribution can maintain, as exemplified by the spatial heterogeneity identified in this study. Both types of distributions can have a fundamental impact on the underlying electronic properties of alloys. The real-space heterogeneity, particularly the domain size of SRO, is also expected to play an important role in other properties, *e.g.*, mechanical and transport properties, of complex alloys.

DATA AVAILABILITY

The trained NEP machine learning potentials are freely available at: https://github.com/shdchen/NEP_GeSn_Alloys_SRO_v1.

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

ACKNOWLEDGEMENTS

The authors acknowledge Zheyong Fan for helpful discussion. This work was supported as part of the μ-ATOMS, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0023412. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0027056. The authors also acknowledge GW High Performance Computing for the computing support.

AUTHOR CONTRIBUTIONS

S.C. and T.L. conceived and supervised the research, S.C., X.J., W.Z., and T.L. prepared the training and testing datasets, S.C. tested various hyperparameters and trained the NEP4 MLP models, S.C. performed band structure calculations, S.C., X.J., and W.Z. performed the validation and analyzed the results. All authors contributed to the writing of the manuscript.

COMPETING INTERESTS

The Authors declare no Competing Financial or Non-Financial Interests.

[1] R. A. Soref and C. H. Perry, Predicted band gap of the new semiconductor SiGeSn, Journal of Applied Physics **69**, 539 (1991).

[2] F. Gencarelli, B. Vincent, J. Demeulemeester, A. Vantomme, A. Moussa, A. Franquet, A. Kumar, H. Bender, J. Meersschaut, W. Vandervorst, R. Loo, M. Caymax, K. Temst, and M. Heyns, Crystalline Properties and Strain Relaxation Mechanism of CVD Grown GeSn, ECS J. Solid State Sci. Technol. **2**, P134 (2013).

[3] S. Wirths, D. Buca, and S. Mantl, Si–Ge–Sn alloys: From growth to applications, Progress in Crystal Growth and Characterization of Materials **62**, 1 (2016).

[4] S. A. Ghetmiri, W. Du, J. Margetis, A. Mosleh, L. Cousar, B. R. Conley, L. Domulevicz, A. Nazzal, G. Sun, R. A. Soref, J. Tolle, B. Li, H. A. Naseem, and S.-Q. Yu, Direct-bandgap GeSn grown on silicon with 2230 nm photoluminescence, Appl. Phys. Lett. **105**, 151109 (2014).

[5] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca, and D. Grützmacher, Lasing in direct-bandgap GeSn alloy grown on Si, Nature Photonics **9**, 88 (2015).

[6] D. Stange, S. Wirths, R. Geiger, C. Schulte-Braucks, B. Marzban, N. von den Driesch, G. Mussler, T. Zabel, T. Stoica, J.-M. Hartmann, S. Mantl, Z. Ikonic, D. Grützmacher, H. Sigg, J. Witzens, and D. Buca, Optically Pumped GeSn Microdisk Lasers on Si, ACS Photonics **3**, 1279 (2016).

[7] V. Reboud, A. Gassenq, N. Pauc, J. Aubin, L. Milord, Q. M. Thai, M. Bertrand, K. Guilloy, D. Rouchon, J. Rothman, T. Zabel, F. Armand Pilon, H. Sigg, A. Chelnokov, J. M. Hartmann, and V. Calvo, Optically pumped GeSn micro-disks with 16% Sn lasing at 3.1 μm up to 180 K, Appl. Phys.

Lett. **111**, 092101 (2017).
[8] S. Al-Kabi, S. A. Ghetmiri, J. Margetis, T. Pham, Y. Zhou, W. Dou, B. Collier, R. Quinde, W. Du, A. Mosleh, J. Liu, G. Sun, R. A. Soref, J. Tolle, B. Li, M. Mortazavi, H. A. Naseem, and S.-Q. Yu, An optically pumped 2.5 μm GeSn laser on Si operating at 110 K, Appl. Phys. Lett. **109**, 171105 (2016).
[9] J. Margetis, S. Al-Kabi, W. Du, W. Dou, Y. Zhou, T. Pham, P. Grant, S. Ghetmiri, A. Mosleh, B. Li, J. Liu, G. Sun, R. Soref, J. Tolle, M. Mortazavi, and S.-Q. Yu, Si-Based GeSn Lasers with Wavelength Coverage of 2–3 μm and Operating Temperatures up to 180 K, ACS Photonics **5**, 827 (2018).
[10] W. Dou, Y. Zhou, J. Margetis, S. A. Ghetmiri, S. Al-Kabi, W. Du, J. Liu, G. Sun, R. A. Soref, J. Tolle, B. Li, M. Mortazavi, and S.-Q. Yu, Optically pumped lasing at 3 μm from compositionally graded GeSn with tin up to 22.3%, Opt. Lett. **43**, 4558 (2018).
[11] M. Kurosawa, M. Kato, T. Yamaha, N. Taoka, O. Nakatsuka, and S. Zaima, Near-infrared light absorption by polycrystalline SiSn alloys grown on insulating layers, Applied Physics Letters **106**, 171908 (2015).
[12] J. Tolle, A. V. G. Chizmeshya, Y. Y. Fang, J. Kouvetakis, V. R. D'Costa, C. W. Hu, J. Menéndez, and I. S. T. Tsong, Low temperature chemical vapor deposition of Si-based compounds via SiH3SiH2SiH3: Metastable SiSn/GeSn/Si(100) heteroepitaxial structures, Applied Physics Letters **89**, 231924 (2006).
[13] T. T. McCarthy, Z. Ju, S. Schaefer, S.-Q. Yu, and Y.-H. Zhang, Momentum(k)-space carrier separation using SiGeSn alloys for photodetector applications, Journal of Applied Physics **130**, 223102 (2021).
[14] O. Moutanabbir, S. Assali, X. Gong, E. O'Reilly, C. A. Broderick, B. Marzban, J. Witzens, W. Du, S.-Q. Yu, A. Chelnokov, D. Buca, and D. Nam, Monolithic infrared silicon photonics: The rise of (Si)GeSn semiconductors, Applied Physics Letters **118**, 110502 (2021).
[15] B. Cao, S. Chen, X. Jin, J. Liu, and T. Li, Short-Range Order in GeSn Alloy, ACS Appl. Mater. Interfaces **12**, 57245 (2020).
[16] X. Jin, S. Chen, C. Lemkan, and T. Li, Role of local atomic short-range order distribution in alloys: Why it matters in Si-Ge-Sn alloys, Phys. Rev. Mater. **7**, L111601 (2023).
[17] X. Jin, S. Chen, and T. Li, Short-range order in SiSn alloy enriched by second-nearest-neighbor repulsion, Physical Review Materials **5**, 104606 (2021).
[18] X. Jin, S. Chen, and T. Li, Coexistence of two types of short-range order in Si–Ge–Sn medium-entropy alloys, Communications Materials **3**, 66 (2022).
[19] S. Liu, A. C. Covian, X. Wang, C. T. Cline, A. Akey, W. Dong, S. Yu, and J. Liu, 3D Nanoscale Mapping of Short-Range Order in GeSn Alloys, Small Methods , 2200029 (2022).
[20] J. Z. Lentz, J. C. Woicik, M. Bergschneider, R. Davis, A. Mehta, K. Cho, and P. C. McIntyre, Local ordering in Ge/Ge–Sn semiconductor alloy core/shell nanowires revealed by extended x-ray absorption fine structure (EXAFS), Applied Physics Letters **122**, 062103 (2023).
[21] A. A. Corley-Wiciak, S. Chen, O. Concepción, M. H. Zoellner, D. Grützmacher, D. Buca, T. Li, G. Capellini, and D. Spirito, Local alloy order in a $Ge_{1-x}Sn_x$/Ge epitaxial layer, Phys. Rev. Appl. **20**, 024021 (2023).
[22] R. Zhang, S. Zhao, J. Ding, Y. Chong, T. Jia, C. Ophus, M. Asta, R. O. Ritchie, and A. M. Minor, Short-range order and its impact on the CrCoNi medium-entropy alloy, Nature **581**, 283 (2020).
[23] J. Behler, Perspective: Machine learning potentials for atomistic simulations, The Journal of Chemical Physics **145**, 170901 (2016).
[24] V. L. Deringer, M. A. Caro, and G. Csányi, Machine Learning Interatomic Potentials as Emerging Tools for Materials Science, Advanced Materials **31**, 1902765 (2019).
[25] T. Mueller, A. Hernandez, and C. Wang, Machine learning for interatomic potential models, The Journal of Chemical Physics **152**, 050902 (2020).
[26] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Machine Learning for Molecular Simulation, Annual Review of Physical Chemistry **71**, 361 (2020).
[27] Y. Mishin, Machine-learning interatomic potentials for materials science, Acta Materialia **214**, 116980 (2021).
[28] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Machine Learning Force Fields, Chemical Reviews **121**, 10142 (2021).
[29] Z. Fan, Z. Zeng, C. Zhang, Y. Wang, K. Song, H. Dong, Y. Chen, and T. Ala-Nissila, Neuroevolution machine learning potentials: Combining high accuracy and low cost in atomistic simulations and application to heat transport, Phys. Rev. B **104**, 104309 (2021).
[30] Z. Fan, Improving the accuracy of the neuroevolution machine learning potential for multi-component systems, Journal of Physics: Condensed Matter **34**, 125902 (2022).
[31] Z. Fan, Y. Wang, P. Ying, K. Song, J. Wang, Y. Wang, Z. Zeng, K. Xu, E. Lindgren, J. M. Rahm, A. J. Gabourie, J. Liu, H. Dong, J. Wu, Y. Chen, Z. Zhong, J. Sun, P. Erhart, Y. Su, and T. Ala-Nissila, GPUMD: A package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations, The Journal of Chemical Physics **157**, 114801 (2022).
[32] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).
[33] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. **104**, 136403 (2010).
[34] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, Journal of Computational Physics **285**, 316 (2015).
[35] A. V. Shapeev, Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials, Multiscale Modeling & Simulation **14**, 1153 (2016).

[36] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B **99**, 014104 (2019).

[37] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nature Communications **13**, 2453 (2022).

[38] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 11423–11436.

[39] Z. Fan, W. Chen, V. Vierimaa, and A. Harju, Efficient molecular dynamics simulations with many-body potentials on graphics processing units, Computer Physics Communications **218**, 10 (2017).

[40] Z. Fan, Y. Xiao, Y. Wang, P. Ying, S. Chen, and H. Dong, Combining linear-scaling quantum transport and machine-learning molecular dynamics to study thermal and electronic transports in complex materials (2023), arXiv:2310.15314 [cond-mat.mtrl-sci].

[41] K. Song, R. Zhao, J. Liu, Y. Wang, E. Lindgren, Y. Wang, S. Chen, K. Xu, T. Liang, P. Ying, N. Xu, Z. Zhao, J. Shi, J. Wang, S. Lyu, Z. Zeng, S. Liang, H. Dong, L. Sun, Y. Chen, Z. Zhang, W. Guo, P. Qian, J. Sun, P. Erhart, T. Ala-Nissila, Y. Su, and Z. Fan, General-purpose machine-learned potential for 16 elemental metals and their alloys (2023), arXiv:2311.04732 [cond-mat.mtrl-sci].

[42] T. J. Rivlin, *Chebyshev Polynomials* (Courier Dover Publications, 2020).

[43] J. Tersoff, New empirical model for the structural properties of silicon, Phys. Rev. Lett. **56**, 632 (1986).

[44] J. Tersoff, Empirical interatomic potential for silicon with improved elastic properties, Phys. Rev. B **38**, 9902 (1988).

[45] T. Schaul, T. Glasmachers, and J. Schmidhuber, High Dimensions and Heavy Tails for Natural Evolution Strategies, in *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11 (Association for Computing Machinery, New York, NY, USA, 2011) pp. 845–852.

[46] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, Natural evolution strategies, Journal of Machine Learning Research **15**, 949 (2014).

[47] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of State Calculations by Fast Computing Machines, J. Chem. Phys. **21**, 1087 (1953).

[48] G. Kresse and J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B **47**, 558 (1993).

[49] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[50] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Computational Materials Science **6**, 15 (1996).

[51] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[52] D. M. Ceperley and B. J. Alder, Ground State of the Electron Gas by a Stochastic Method, Phys. Rev. Lett. **45**, 566 (1980).

[53] C. Eckhardt, K. Hummer, and G. Kresse, Indirect-to-direct gap transition in strained and unstrained sn_xge_{1-x} alloys, Phys. Rev. B **89**, 165201 (2014).

[54] M. P. Polak, P. Scharoch, and R. Kudrawiec, The electronic band structure of $Ge_{1-x}Sn_x$ in the full composition range: Indirect, direct, and inverted gaps regimes, band offsets, and the Burstein–Moss effect, J. Phys. D: Appl. Phys. **50**, 195103 (2017).

[55] P. Haas, F. Tran, and P. Blaha, Calculation of the lattice constant of solids with semilocal functionals, Phys. Rev. B **79**, 085104 (2009).

[56] F. Tran and P. Blaha, Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential, Phys. Rev. Lett. **102**, 226401 (2009).

[57] H. J. Monkhorst and J. D. Pack, Special points for Brillouin-zone integrations, Physical Review B **13**, 5188 (1976).

[58] J. M. Cowley, An approximate theory of order in alloys, Phys. Rev. **77**, 669 (1950).

[59] V. Popescu and A. Zunger, Effective Band Structure of Random Alloys, Phys. Rev. Lett. **104**, 236403 (2010).

[60] O. Rubel, A. Bokhanchuk, S. J. Ahmed, and E. Assmann, Unfolding the band structure of disordered solids: From bound states to high-mobility Kane fermions, Phys. Rev. B **90**, 115202 (2014).

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).

[62] A. Tamm, A. Aabloo, M. Klintenberg, M. Stocks, and A. Caro, Atomic-scale properties of Ni-based FCC ternary, and quaternary alloys, Acta Materialia **99**, 307 (2015).

[63] J. Ding, Q. Yu, M. Asta, and R. O. Ritchie, Tunable stacking fault energies by tailoring local chemical order in CrCoNi medium-entropy alloys, PNAS **115**, 8919 (2018).

[64] S. Chen, Z. H. Aitken, S. Pattamatta, Z. Wu, Z. G. Yu, D. J. Srolovitz, P. K. Liaw, and Y.-W. Zhang, Simultaneously enhancing the ultimate strength and ductility of high-entropy alloys via short-range ordering, Nature Communications **12**, 4953 (2021).

[65] A. Mascarenhas, *Spontaneous Ordering in Semiconductor Alloys* (Springer Science & Business Media, 2002).

[66] S. Chen, X. Jin, W. Zhao, and T. Li, unpublished.

Supporting Information: Intricate short-range order in GeSn alloys revealed by atomistic simulations with highly accurate and efficient machine-learning potentials

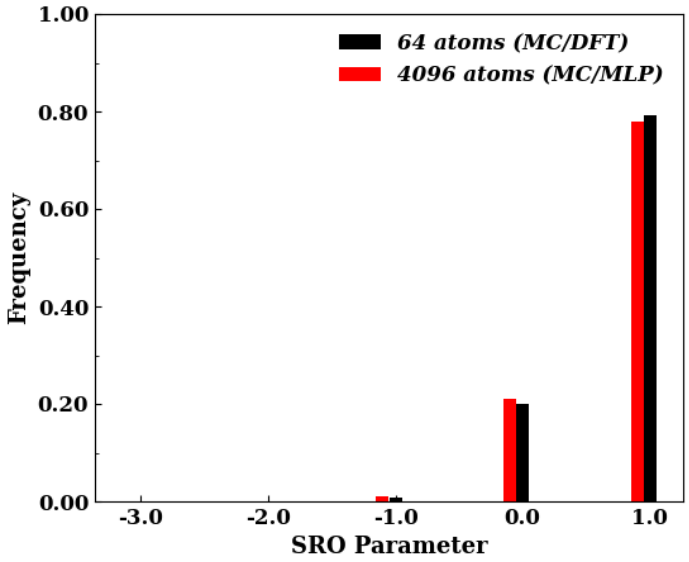

FIG. 6. Distribution of local atomic SRO parameters of 1NN Sn-Sn in $Ge_{0.75}Sn_{0.25}$ alloys, obtained from MC/DFT using 64-atom supercell and MC/MLP using 4096-atom supercell, showing that MC/MLP using larger supercell with our developed MLP can reproduce the same average 1NN Sn-Sn SRO parameters and their distribution as those from MC/DFT using smaller supercell.

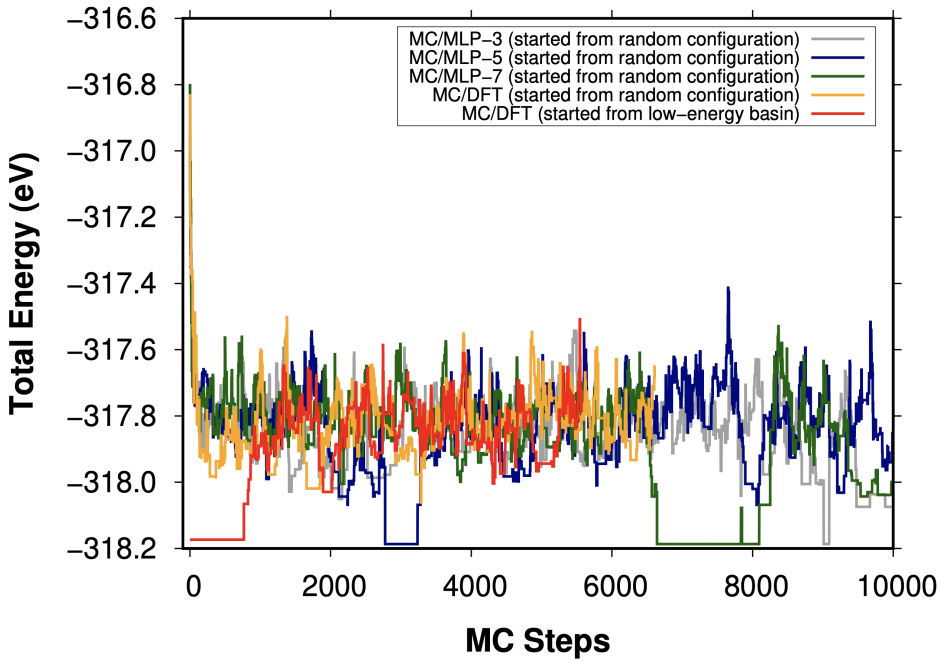

FIG. 7. MC/DFT confirmation of the coexistence of two energy basins. The MC/DFT sampling trajectory (in red) initiates from the structure within the low-energy basin, remains in the low-energy basin for a considerable number of MC steps, consistent with the MC/MLP samplings (in gray, blue, green). Subsequently, it transitions to a high-energy basin, aligning with previous MC/DFT sampling [15] (in orange) and MC/MLP samplings (in gray, blue, green).

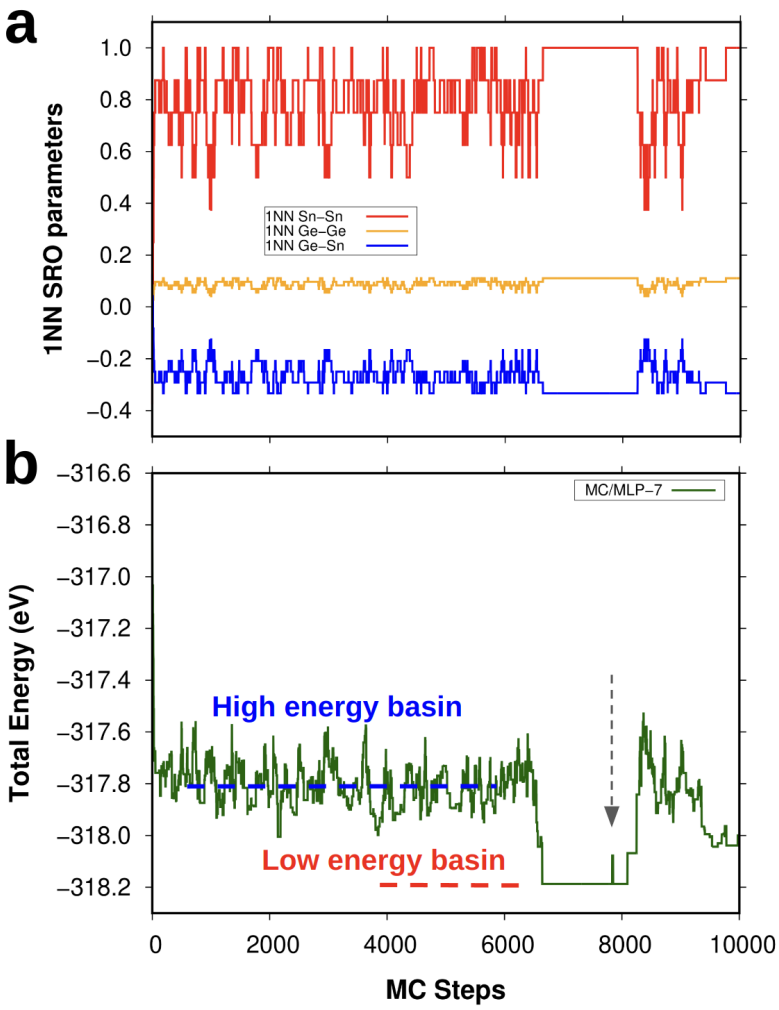

FIG. 8. (**a**) Variation in SRO parameters of 1NN Sn-Sn (α^1_{SnSn}, red), 1NN Ge-Ge (α^1_{GeGe}, orange), and 1NN Ge-Sn (α^1_{GeSn}, blue) as a function of MC steps in MC/MLP-7 trajectory. (**b**)The corresponding energy variation as a function of MC steps in MC/MLP-7 trajectory. A slight jump in the low energy basin is observed. Interestingly, this small energy variation is not captured by the 1NN Sn-Sn SRO parameters, nor the 1NN Ge-Ge, or 1NN Ge-Sn SRO parameters. Instead, the subtle total energy change is found well captured when including the 2NN and 3NN Sn-Sn SRO parameters, as shown in Fig. 2**b**.