ARTICLE OPEN

Metadynamics sampling in atomic environment space for collecting training data for machine learning potentials

Dongsun Yoo [1,2,3], Jisu Jung [1,3], Wonseok Jeong[1] and Seungwu Han [1✉]

The universal mathematical form of machine-learning potentials (MLPs) shifts the core of development of interatomic potentials to collecting proper training data. Ideally, the training set should encompass diverse local atomic environments but conventional approaches are prone to sampling similar configurations repeatedly, mainly due to the Boltzmann statistics. As such, practitioners handpick a large pool of distinct configurations manually, stretching the development period significantly. To overcome this hurdle, methods are being proposed that automatically generate training data. Herein, we suggest a sampling method optimized for gathering diverse yet relevant configurations semi-automatically. This is achieved by applying the metadynamics with the descriptor for the local atomic environment as a collective variable. As a result, the simulation is automatically steered toward unvisited local environment space such that each atom experiences diverse chemical environments without redundancy. We apply the proposed metadynamics sampling to H:Pt(111), GeTe, and Si systems. Throughout these examples, a small number of metadynamics trajectories can provide reference structures necessary for training high-fidelity MLPs. By proposing a semi-automatic sampling method tuned for MLPs, the present work paves the way to wider applications of MLPs to many challenging applications.

INTRODUCTION

By delivering the accuracy of density-functional theory (DFT) calculations at much lower costs, atomistic simulations based on machine-learning potentials (MLPs) are being established as a new pillar in computational material science[1]. Most MLPs utilize the locality of quantum systems and so the computational cost increases linearly with respect to the system size, which is a significant advantage over DFT with a cubic scaling[2]. Until now, various types of MLPs have been proposed; neural network potential (NNP)[3], Gaussian approximation potential (GAP)[4], moment tensor potential[5], deep tensor neural network[6], and gradient-domain machine learning[7]. In particular, the NNP and GAP are garnering wide interests with applications to challenging simulations such as crystallization behaviors of GeTe[8,9] and $Ge_2Sb_2Te_5$[10], Ni-silicidation process[11], proton transfer at the ZnO–water interface[12], structure search of $Pt_{13}H_x$ clusters[13], crystal structure prediction[14], and identification of active sites in bimetallic catalysts for CO_2 reduction[15].

At the heart of the traditional classical potential is the mathematical formula that captures underlying bonding natures. In contrast, universal mathematical structures of MLPs shift the core of potential development to collecting a proper training set that defines atomic environments wherein the trained MLP is valid. Ideally, the training set should encompass diverse local configurations that may appear in target simulations. In usual practices, the training set is selected based on crystal-derived structures and their molecular dynamics (MD) trajectories. However, MD simulations are conditioned by the Boltzmann statistics, which over-represents low-energy regions and can sample only a few distinct configurations separated by low thermal barriers. As a result, a large pool of reference structures is handpicked manually, which demands expertise from practitioners as well as several iterative refinements of MLPs[16–18].

We note that methods such as active learning[19–21], random structure search[22,23], and entropy-maximization[24] focus on sampling diverse configurations in automatic fashion, but they were aimed for specific purposes or have not been employed in complicated simulations as far as we are aware.

The above discussions call for a sampling method specifically aiming to prepare training sets for MLPs, which is tuned to collect local atomic environments as diverse as possible within the time- and size-scale of DFT calculations. In addition, the sampled configurations should be relevant in the intended simulations. We herein propose one such approach based on metadynamics[25]. The metadynamics defies the Boltzmann distribution by accumulating bias potentials along the collective variables (CVs). Instead of usual implementations that formulate CVs from a set of atomic positions in the real space[26], we employ as CVs the coordinates in the abstract atomic-environment space, which is spanned by the atom-centered symmetry-function vector (**G**)[27]. Being widely used as input features of NNPs, the **G** vector parametrizes local atomic environments into fixed-length vectors via integrating radial and angular distributions of neighboring atoms. By accumulating bias potentials in the **G** space, the present metadynamics (abbreviated as G-metaD hereafter) drives each atom to evolve towards unvisited points in the **G** space. In addition, G-metaD is controlled by a few hyperparameters and can start with simple initial structures, requiring less expertise than the conventional MD-based sampling. To note, Bonati and Parrinello also used metadynamics to generate a training set for MLP[28]. In this case, however, the CV was designed to sample a particular process of interest (nucleation events), whereas the present approach is generally applicable to any target simulations and aims to sample diverse configurations. In another publication, Herr et al. suggested a metadynamics sampling method using the distance matrix of the whole system as a CV, which enhanced the stability

[1]Department of Materials Science and Engineering and Research Institute of Advanced Materials, Seoul National University, Seoul, Korea. [2]Present address: Samsung Display Co., Ltd., 1, Samsung-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Korea. [3]These authors contributed equally: Dongsun Yoo, Jisu Jung. ✉email: hansw@snu.ac.kr

of MD simulations compared to conventional MD sampling[29]. However, individual atoms in this approach can still repeatedly encounter similar local environments because the CV is based on the total system configuration.

In the following section, we formulate the G-metaD and demonstrate its applications to three systems: H:Pt(111), GeTe, and Si. The first model, H:Pt(111), is chosen to directly compare the sampling style between the conventional MD and G-metaD. The other two examples, GeTe and Si, were studied by using the NNP or GAP[8,9,17,30,31]. We choose these materials to benchmark the performance of NNPs trained by G-metaD trajectories against the state-of-the-art MLPs trained over a large number of structures that were prepared manually.

RESULTS

Metadynamics simulation

The present G-metaD employs the $\mathbf{G}$ vector as the CV. The local bias potential (u_b) is defined as a function of $\mathbf{G}$, and the summation of atomic local biases constitutes the total bias potential (U_b) applied on the system:

$$U_b(\{\mathbf{R}(t)\}) = \sum_{i=1}^{N_{at}} u_b(\mathbf{G}_i(t)),$$ (1)

where $\{\mathbf{R}(t)\}$ and $\mathbf{G}_i(t)$ are the set of all position vectors and the symmetry-function vector of the ith atom at time t, respectively, and N_{at} is the number of atoms in the system. The biasing force on each atom is computed as follows:

$$F_{i,a} = -\frac{\partial U_b}{\partial R_{i,a}} = -\sum_{j=1}^{N_{at}} \sum_{s=1}^{N_G} \frac{\partial U_b}{\partial G_{j,s}} \frac{\partial G_{j,s}}{\partial R_{i,a}},$$ (2)

where $F_{i,a}$ and $R_{i,a}$ are the a-component ($a = x$, y, and z) of the force and position vectors of the ith atom, respectively, and $G_{j,s}$ is the sth component of $\mathbf{G}_j$ with the dimension of N_G. For multi-component systems, u_b is defined independently for each atomic species.

We construct the local bias u_b in Eq. (1) using Gaussians centered at $\mathbf{G}$ points visited by each atom. Since the elements of $\mathbf{G}$ vectors are highly correlated to each other, it is ineffective to adopt isotropic Gaussians with a fixed width. This is illustrated in Fig. 1a, which schematically shows a typical distribution of training points (gray dots) along two components (G_i and G_j). To sample the distribution with isotropic Gaussian biases (black dots with circles whose radius means the Gaussian width), many Gaussians should be accumulated because of the highly anisotropic distribution. To overcome this problem, we employ
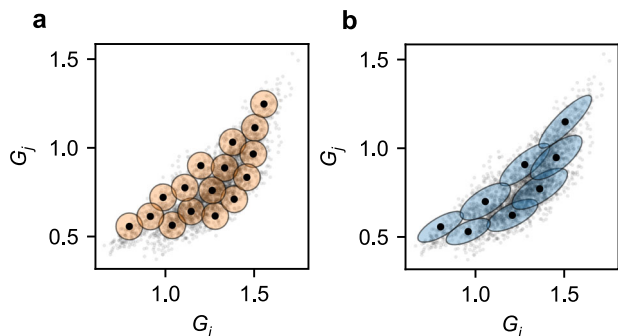
geometry-adapted Gaussians, which was developed to reconstruct the accurate potential energy surface (PES) from metaD by adjusting the shape and size of Gaussian biases according to the distribution of visited CVs[32]. The present G-metaD reformulates this approach and accumulates adaptive Gaussians centered at visited $\mathbf{G}$ points as follows:

$$u_b(\mathbf{G}(t)) = h \int_0^t \sum_{i=1}^{N_{at}} \sum_{n \in \mathbb{N}} \exp\left[-\frac{1}{2}(\mathbf{G}(t) - \mathbf{G}_i(t'))^T \sum(\mathbf{G}_i(t'))^{-1}(\mathbf{G}(t) - \mathbf{G}_i(t')) \right] \delta(t' - n\tau) dt'.$$ (3)

The covariance matrix Σ in Eq. (3) is given by:

$$\sum_{jk}(\mathbf{G}) = \sigma^2 \sum_{i=1}^{N_{at}} \sum_{a=x,y,z} \frac{\partial G_j}{\partial R_{i,a}} \frac{\partial G_k}{\partial R_{i,a}} + \varepsilon \delta_{jk},$$ (4)

where G_j and G_k are the jth and kth components of $\mathbf{G}$, respectively. In Eqs. (3) and (4), hyperparameters h, σ, and τ represent the height and width of Gaussian potentials and time interval of bias updates, respectively. The high correlations among components of $\mathbf{G}$ render Σ^{-1} to be numerically unstable. To prevent the divergence, a small regularization term ε (fixed to 10^{-4}) is added to the diagonal components in Eq. (4). According to Eqs. (3) and (4), the width of the Gaussian bias is adjusted anisotropically such that the Gaussian shape resembles the data distribution as shown in Fig. 1b. This enables the G-metaD to search the relevant regions with much smaller number of bias potentials than in Fig. 1a.

There are three hyperparameters in G-metaD: h, σ, and τ. Being related to the bias strength, h and σ control the height and width of Gaussian potentials, respectively. In the case of h, a proper value is chosen such that the magnitude of U_b has an order similar to the thermal energy and the simulation remains to be stable. For σ, if it is too small, a large number of bias potentials would be needed to fill a basin of the PES. In contrast, a large σ can obscure the curvature of the PES, thereby causing under-sampling. We find that σ around 1 Å is a reasonable choice. Lastly, τ should be long enough for the system to respond to the updated bias force, and short enough that the metadynamics trajectories can search diverse configurations in the limited simulation time. Our experiences indicate that 20 fs is a sound choice for τ, which is used throughout the present work. Note that these hyperparameters could be assigned differently for each type of atom.

In general, high dimensional variables such as $\mathbf{G}$ vectors are not appropriate as CVs in metadynamics. However, unlike conventional CVs, components of $\mathbf{G}$ vectors are highly correlated with each other. With the regularization term (ε) in Eq. (4), this effectively limits the dimensions being explored. In detail, due to the high correlations, the magnitude of principal components decays rapidly such that eigenvalues of the covariance matrix become far smaller than ε after a few principal components. This is equivalent to applying large and wide biases along those dimensions, and the exploration is effectively limited to the first few principal components while the other components are ignored. This renders the present metadynamics sensible despite the high dimension of CVs.

We implement a pair style that computes the bias potential into the LAMMPS package[33] using the SIMPLE-NN library[34]. By operating the client-server mode, one can interface LAMMPS with other ab initio codes such as VASP[35] to perform the G-metaD.

Hydrogen on the Pt(111) surface

To compare G-metaD and conventional MD in terms of the sampling style, we investigate diffusion of the H atom on the Pt (111) surface. Noble metals are widely used as efficient catalysts for H-involved reactions such as hydrogen evolution reactions[36] and CO_2 reduction[37]. An accurate description of the H diffusion on the metal surface is important for simulating these reactions. Furthermore, it has been reported that H atoms can diffuse into
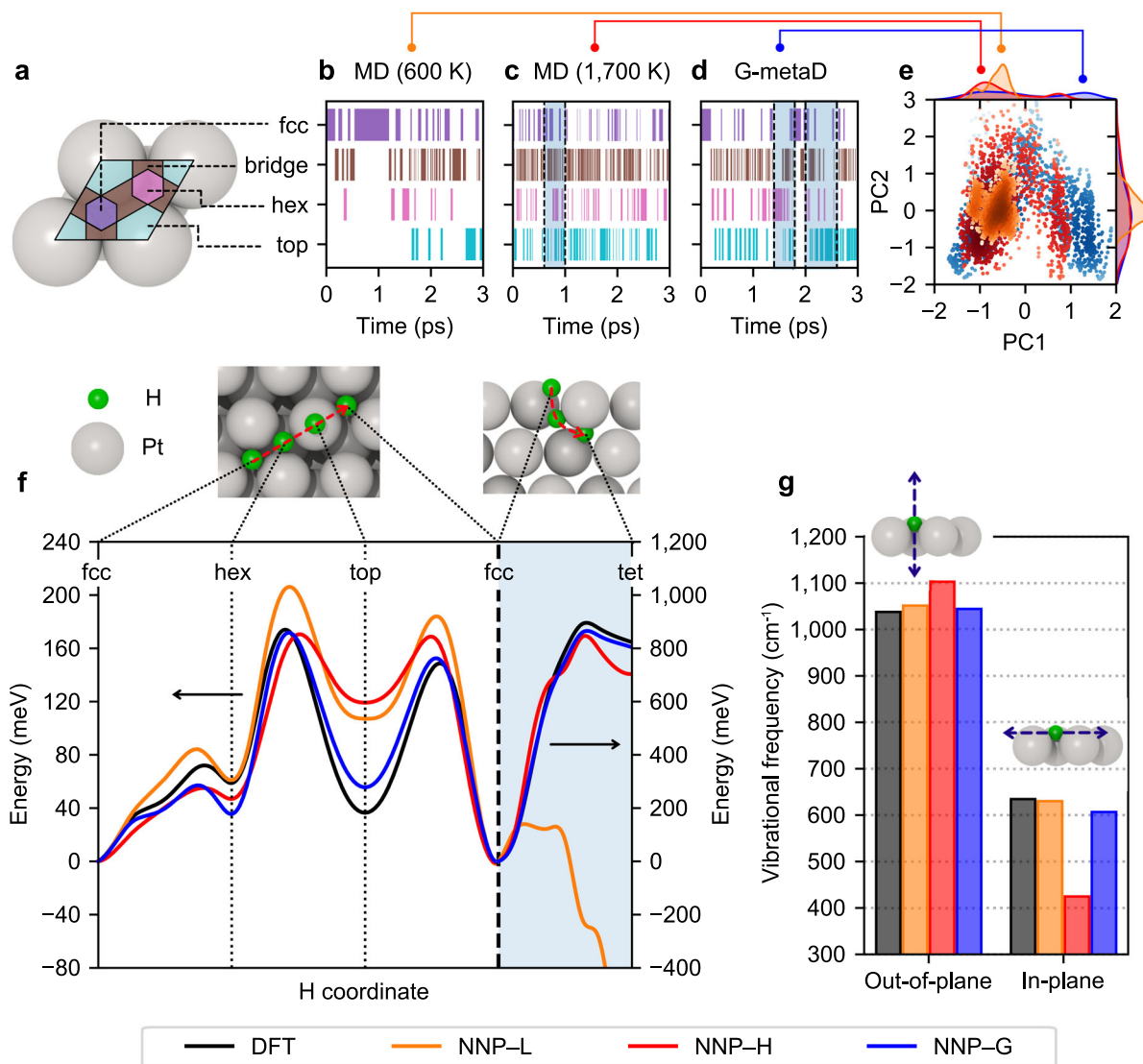

Fig. 1 The schematic description of bias potential shape in the G space projected on certain components G_i and G_j. a Isotropic Gaussian potential with a fixed width. **b** Adaptive Gaussian potential. Gray dots represent the $\mathbf{G}$ vector points sampled during the MD simulation. Black dots represent the $\mathbf{G}$ vector points where bias potential is evaluated and semi-transparent circles and ovals represent bias potentials centered on the black dots.

Fig. 2 **Sampled configurations and performances of the corresponding NNPs. a** Characteristic area on Pt(111) surface. **b–d** The classification of H sites along the time for MD at 600 K, 1700 K, and G-metaD at 600 K, respectively. The shaded area means that the H atom stays in the sublayer. **e** Distribution of sampled points of H atom on principal components axes. **f** The minimum energy paths along fcc → hex → top → fcc → tet (subsurface tetrahedral site). **g** Two vibrational frequencies of the H atom at the fcc site. NNP-L, NNP-H, and NNP-G stand for the NNP trained over trajectories from MD at 600 K, 1700 K, and G-metaD at 600 K, respectively. (The atomic configurations are visualized by OVITO[62]).

the subsurface, influencing the total diffusion kinetics and reaction rates[38,39]. Thus, sampling various H sites on the surface as well as in the subsurface would be important for training MLPs that aim to simulate catalytic reactions.

To obtain the training data, we carry out three simulations on (3 × 3)-Pt(111) with one H atom adsorbed on the surface; standard MD at 600 and 1700 K and G-metaD at 600 K under NVT conditions. In the case of G-metaD, the bias potential is applied only on the H atom with h and σ of 24 meV and 1.0 Å, respectively. The total simulation time is 3 ps for every simulation. The detailed set up for DFT calculations are presented in the "Methods" section and simulation movies of MDs at 600 K (Supplementary Video 1), 1700 K (Supplementary Video 2), and G-metaD at 600 K (Supplementary Video 3) are also provided. At lower temperatures than 1700 K, say, 1500 K, the diffusion into subsurface is not observed during 3-ps MD simulations. In Fig. 2a, the trajectories of the H atom are classified into four regions; face-centered cubic (fcc), hexagonal (hex), bridge, and top sites. At 0 K, the lowest energies

within each region are 0 (fcc; the reference site), 59 (hex), 47 (bridge), and 37 (top) meV. The lowest energy site in the sublayer is the tetrahedral site right below the top site with the energy of 823 meV in reference to the fcc site.

The temporal evolution of visited sites are displayed in Fig. 2b–d. At 600 K, the H atom stays mostly at the fcc site that is the lowest in the potential energy, which is consistent with the Boltzmann distribution. In addition, the H atom does not penetrate into the subsurface due to a high diffusion barrier of 0.9 eV (see below). At the elevated temperature of 1700 K in Fig. 2c, various sites are sampled more or less evenly and subsurface diffusion is observed (shaded area). On the other hand, G-metaD at 600 K also samples various sites including the sublayer (see Fig. 2d). The H atom in the G-metaD stays within the sublayer for ~1 ps out of 3-ps simulation time, in contrast to ~0.5 ps duration in 1700-K MD (see Fig. 2c). Figure 2e shows the distribution of the three trajectories projected onto major principal axes from principal component analysis. It is seen that the G-metaD covers a wider area than 600- or 1700-K MD.

Next, we train three NNPs by employing trajectories from each simulation as training data. (The trajectories are sampled every 10 fs.) They are named NNP-L, NNP-H, and NNP-G according to the MD types (600-K MD, 1700-K MD, and G-metaD, respectively). The training and validation errors are similar among the three NNPs and root-mean-squared errors (RMSEs) for energy and force are less than 3 meV/atom and 0.2 eV/Å, respectively. (We refer to the Methods section for the details of NNPs and training procedures.) To compare accuracy of the trained NNPs, we compute in Fig. 2f the minimum energy paths (MEPs) between the three symmetric sites (fcc, hex, and top) using the nudged-elastic-band method with 9 replicas between the symmetric sites[40,41]. The MEP into the subsurface tetrahedral site is also calculated on the right side. For reference, DFT results are also presented, which agree well with literature[42,43]. Overall, the results obtained with the NNP-G best agree with the DFT results except that NNP-G incorrectly estimates the hexagonal site to be more stable than the top site by 19 meV while it is less stable by 22 meV in DFT. The magnitude of the error may appear to be inconsistent with RMSE of 3 meV/atom (see above). However, it is noted that the hexagonal-site energy is computed by the difference in total energies, while RMSE is measured as the total energy difference per atom. (The supercell used in calculating the site energy contains 37 atoms.)

In Fig. 2f, it is notable that both NNP-L and NNP-H show large errors of ~0.1 eV at the top site. The undersampling in 600-K MD for this site (see Fig. 2b) would be responsible for the error with NNP-L. Even though top sites are well sampled in MD at 1700 K (see Fig. 2c), the large error implies that the trajectory fails to capture the low-energy surface because the MD is heavily influenced by wide atomic vibrations. The same reasons account for the errors along the diffusion into the subsurface (shaded area of Fig. 2f); a dramatic failure of NNP-L certainly originates from the absence of data in this region (see Fig. 2b), which is partly resolved by MD at 1700 K. However, a substantial error of ~0.1 eV remains at the subsurface tetrahedral site. This implies that the high-temperature sampling, while useful for overcoming energy barriers, risks undersampling atomic environments around local minima due to the wide vibrations and entropic effects. In contrast, the G-metaD is performed at moderate temperatures so it does not suffer from such problems.

Figure 2g compares vibrational frequencies at the stable fcc sites. There are two independent vibrational modes, out-of-plane and in-plane (twofold). The accuracy with respect to DFT results follows the order of NNP-L > NNP-G > NNP-H. This is consistent with the above observations: the 600-K MD most densely samples the fcc site, resulting in the highest accuracy. In contrast, 1700-K MD undersamples atomic environments near local minima because of large thermal energies. The reasonable accuracy with the NNP-G implies that G-metaD samples enough points around the minima like 600-K MD until the bias fills the basin. (This is confirmed by the principal component analysis around the fcc site (not shown).)

Amorphization of GeTe

The present G-metaD is controlled by two hyperparameters, h and σ. By tuning these two parameters, one can steer the system to explore different regions in the **G**-space, which enables a semi-automatic sampling. We demonstrate this with GeTe, an archetypal phase-change material that has been extensively studied for non-volatile memory devices[44]. Several studies employed NNPs in simulating the amorphous structures and crystallization behaviors of GeTe[8,30,31]. To sample diverse local orders, the training set included liquids, crystals, amorphous phases as well as quenching trajectories. To improve the stability of the simulation, non-stoichiometric phases were also considered[2,8,30]. Here we attempt to prepare the training set for simulating GeTe by utilizing G-metaD only.
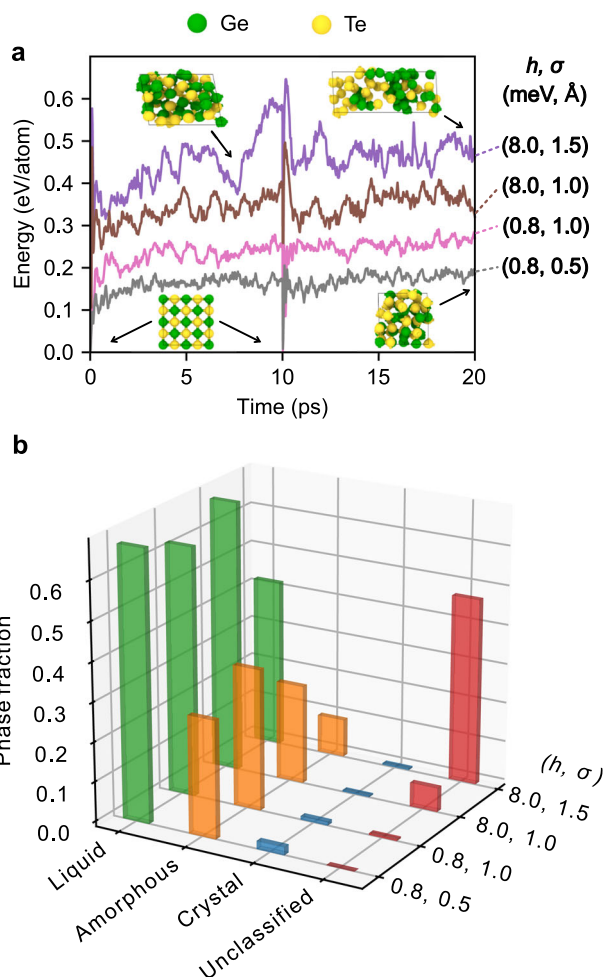

Fig. 3 Energy and fraction of phase in the four G-metaD. a Time evolution of potential energy of four G-metaD along the time. The energy is referenced to that of the rock-salt structure. Structures during G-metaD are shown as inset figures. **b** The phase fractions classified by Mahalanobis distances are shown for each (h, σ).

Starting from the crystalline rock-salt structure, the G-metaD for GeTe is carried out for 20 ps under NPT conditions of 64 atoms, 0 kbar, and 600 K. For simulating the whole melt-quench process as well as crystallization behaviors, it is necessary to sample both liquid structures with high energies and amorphous structures with the local order similar to those in the crystalline phases. To this end, we generate four G-metaD trajectories with different choices of (h, σ): (8.0, 1.5), (8.0, 1.0), (0.8, 1.0), and (0.8, 0.5) in (meV, Å). The movies for four G-metaD trajectories are provided Supplementary Videos 4–7 in the same order of (h, σ). In Fig. 3a, the evolution of potential energies during 20-ps G-metaD is shown for each (h, σ). To sample local orders close to the crystalline structure, we restart G-metaD at 10 ps from the rock-salt structure while maintaining the bias potential accumulated during the preceding 10 ps to avoid redundancy. Here we use h values of 8 or 0.8 meV, which are far smaller than 24 meV in the previous example. This is because there are 32 atoms that contribute to the bias potential (see Eq. (3)) while there was only one atom (H) in the previous case. It is seen in Fig. 3a that at the strongest bias ($h = 8$ meV and $\sigma = 1.5$ Å), the system widely changes and even phase separations are noticeable near the end of the G-metaD (see inset figures at the top). In ref. [8], diffusional mixing of liquid Ge and Te was considered to prevent unphysical phase separations from the ad hoc energy mapping. Such atomic environments are automatically sampled in the

strongly biased G-metaD. In contrast, under the weakest bias strengths ($h = 0.8$ meV and $\sigma = 0.5$ Å), the trajectory remains relatively close to the crystalline structures and appears to mainly sample amorphous-like structures.

To analyze characteristic structures that each G-metaD samples, we introduce the Mahalanobis distance, which is used in measuring distances between a point and distributions[45,46]. (See the "Methods" section for details.) Using the Mahalanobis distance, we classify atomic environments into crystal, amorphous, and liquid structures. If the distance does not satisfy the given criteria for any of the three phases, the sampled **G** point remains to be unclassified. Histograms in Fig. 3b display the phase fractions sampled for each (h, σ). At the strongest bias, unclassified structures are the most dominant. The rapidly accumulating bias potentials drive the system to evolve towards high-energy structures resembling surfaces or unmixed phases. As the bias strength is reduced, the relative portions of bulk structures, in particular rock-salt structures, increases. This analysis indicates that the system can explore distinct regions in the **G** space by tuning the hyperparameters.

Using the four G-metaD trajectories, we train an NNP with the energy, force, and stress RMSEs of 6 meV/atom, 0.24 eV/Å, and 5 kbar, respectively. (Trajectories are sampled every 20 fs.) In Fig. 4a, the equation of states (EOS) for rock-salt ($Fm3m$) and rhombohedral ($R3m$) phases are compared between NNP and DFT. The equilibrium volume and bulk moduli agree with DFT within 1%. Even though the EOS and deformed crystals were not explicitly included in the training set, good agreements are found with both phases. This indicates that G-metaD can automatically sample various lattice distortions around the equilibrium structure. However, the small energy difference between rock-salt and rhombohedral phases (8 meV/atom) is neglected by the NNP (<1 meV/atom).

Next, we perform melt-quench simulations with the NNP and characterize structural properties of resulting liquid and amorphous structures. To compare with DFT on an equal footing, we select a 96-atom supercell for the simulation. The temperature protocol is identical to that in ref. [8]. During the melt-quench process, we do not observe any artefacts such as phase separations. Figures 4b and c compare the total and element-resolved radial-distribution functions (RDFs) for liquid and amorphous phases. Overall, good agreements with DFT are found, which is comparable to the previous studies[8]. In Fig. 4d, we analyze the ring statistics by using the R.I.N.G.S. code[47]. Albeit overestimated in densities, the overall ring distributions including the portion of ABAB-type four-membered rings are similar between NNP and DFT.

We also simulate the crystallization behavior of a 4096-atom supercell at 500 K with an atomic density of the amorphous phase. (See Fig. 4e.) The crystallization behavior is similar to those in refs. [8] and [9]. For a quantitative comparison, we calculate the crystal growth velocity by following the method in refs. [9] and [48], and it is found to be 1.88 m/s at 500 K, in reasonable agreement with 1.89 and 0.52 m/s at the same temperature in refs. [8] and [9], respectively. In ref. [8], it was found that NNP tends to produce flat four-fold rings, resulting in unphysically fast crystallization, which was improved when relaxation paths from the flat to puckered four-fold rings were included in the training set. The present NNP produces flatness in between the conventional and refined NNPs (not shown), implying that the fine details of medium-range order are not well captured by either MD or G-metaD.

General-purpose potential for Si

Due to the limited transferability, most MLPs are trained for specific applications. Developing general-purpose MLPs is a formidable task involving the construction of a huge data set
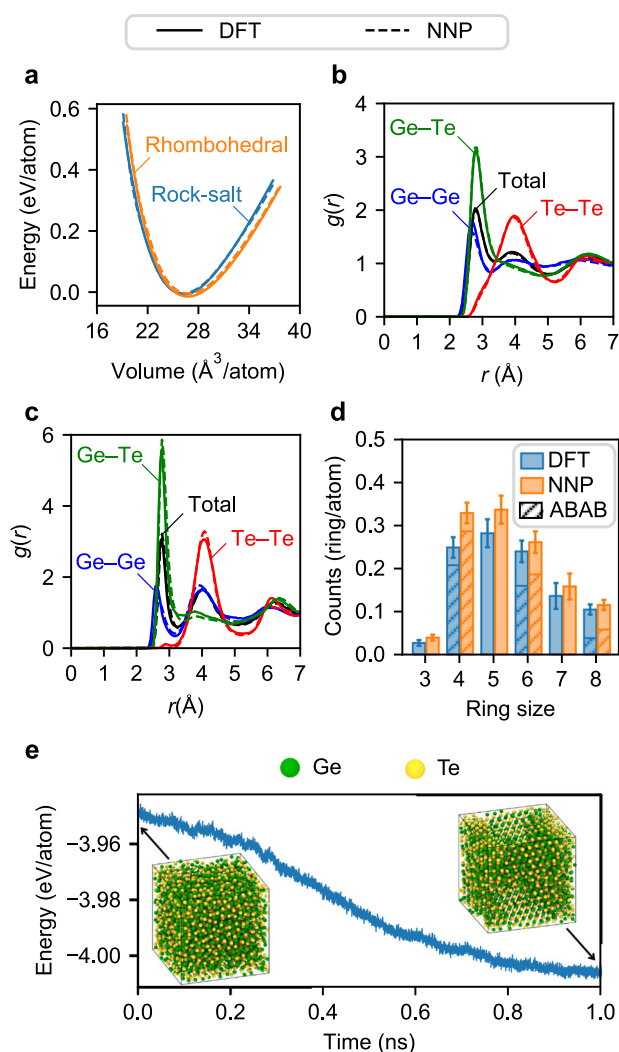

Fig. 4 **Structural properties and crystallization behavior of GeTe obtained by the NNP. a** The energy-volume relation for rock-salt and rhombohedral phases. The energy is referenced to that of the rock-salt phase at the equilibrium. **b**, **c** The total and element-resolved radial-distribution functions ($g(r)$) for liquid at 1000 K (**b**) and amorphous structures at 300 K (**c**). In **a–c** solid and dashed lines indicate DFT and NNP results, respectively. **d** Ring statistics of amorphous GeTe at 300 K. The error bars indicate one standard deviation obtained from four independent MD simulations. The ABAB-type rings (A = Ge and B = Te) for even-membered rings are also shown. **e** The time evolution of potential energy during the crystallization. Initial and final structures are shown as inset figures.

that contains a vast range of chemical environments, which in turn requires deep understanding of the system and possibly several iterations of refinements of MLPs. There have been a few attempts to generate general-purpose MLPs with manually selected data set[17,18]. For example, in ref. [17], a general-purpose GAP potential was developed for Si with the training set covering a long list of distinct structures such as several polymorphs, extended and point defects, slab models, amorphous and liquid phases (29 in total).

Here we demonstrate with Si that the G-metaD can be applied to preparing a training set for the general-purpose potential in convenient ways. Using two sets of hyperparameters ([h (meV), σ (Å)] = [0.4, 1.0] and [0.04, 1.0]), we generate two 30-ps G-metaD trajectories starting from a 64-atom
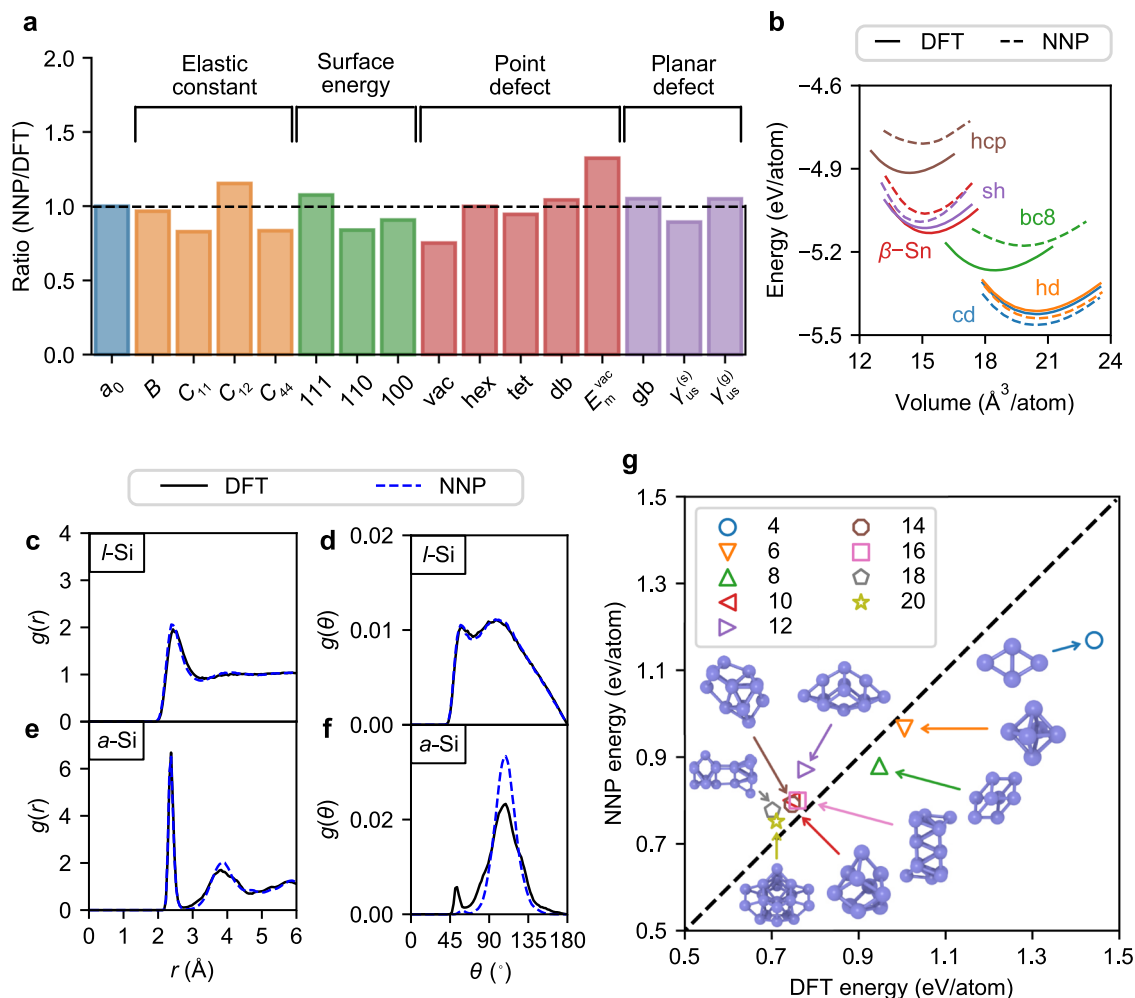
Fig. 5 Comparison between NNP and DFT results for various properties of Si. a Ratios of NNP to DFT for static properties of Si in the diamond structure. Surface energies are calculated for (100)–(2 × 2)[63], (110)–(1 × 1)[64], and (111)–(3 × 3) reconstructions[65]. Defect formation energies for the vacancy (vac) and interstitials (hexagonal (hex), tetrahedral (tet), and dumbbell (db)). E_m^{vac} is the migration energy of the vacancy. For extended defects, gb means (112)Σ3 grain boundary. $\gamma_{us}^{(s)}$ and $\gamma_{us}^{(g)}$ are unstable stacking-fault energies on shuffle and glide planes of the diamond (111) plane, respectively. **b** Equation of states for polymorphs. The abbreviations cd, hd, bc8, and sh stand for cubic diamond, hexagonal diamond, body-centered cubic, and simple hexagonal, respectively. **c** RDF and **d** angular distribution function (ADF; $g(\theta)$ of liquid Si (l-Si) at 2500 K. **e** RDF and **f** ADF of amorphous Si (a-Si) at 300 K. In **b**–**f** solid and dashed lines indicate the reference DFT and NNP results, respectively. **g** Energies of even-number nanoclusters in reference to the equilibrium diamond structure obtained by each method. Insets are structures relaxed by the NNP.

supercell in the cubic-diamond (cd) structure, under the NPT condition (0 kbar and 600 K). Like in GeTe, G-metaD restarts from the crystalline Si every 10 ps. Here we use smaller h values than the previous GeTe example because of the larger number of atoms contributing to the bias (64 vs. 32) and smaller degrees of configurational freedom (unary vs. binary). The simulation time is extended to 30 ps to sufficiently sample local environments far from the diamond structure such as high-temperature liquids and under-coordinated atoms. The movies for G-metaD trajectories for the former (Supplementary Video 8) and latter (Supplementary Video 9) hyperparameters are provided. Using the two G-metaD trajectories sampled every 20 fs, we train an NNP with the energy, force, and stress RMSEs of 23 meV/atom, 0.36 eV/Å, and 5 kbar, respectively.

Figure 5a lists elastic, surface, and defect properties of Si computed with the NNP, which are scaled by the DFT values. These properties were selected by ref. [17]. in benchmarking the general-purpose potential of Si. Overall, NNP results show reasonable agreements with DFT although the test structures

were not explicitly included in the training set. The mean absolute error for these properties is 11%, higher than 6% in ref. [17]. where the data set was constructed manually. To note, planar defects were not included in the training data of ref. [17], and errors of the unstable stacking fault energies for shuffle ($\gamma_{us}^{(s)}$) and glide ($\gamma_{us}^{(g)}$) were −16% and 13%, respectively, which are larger than −11% and 5% in the present work.

Figure 5b shows EOS of various phases of Si. We note significant deviations for high-pressure phases such as hcp and bc8. During G-metaD simulations, atoms in the same supercell are driven to different chemical environments (i.e., **G** vectors), unlike crystalline structures where atoms share a few **G** vectors. Therefore, the prediction error tends to increase in crystalline structures with local orders different from the initial one. This can be improved by augmenting the training set (see the Discussion section). On the other hand, the structural properties of liquid and amorphous phases in Figs. 5c–f are in reasonable agreement with DFT except that the tetrahedral unit in the amorphous Si (a-Si) is more rigid with the NNP (see Fig. 5f).

Lastly, we calculate in Fig. 5g energies of Si nanoclusters with the number of atoms ranging over 4–20[49–51]. We adopt geometries from refs. [49] and [50]. and relax them using DFT and NNP. (The structures in the inset are obtained by NNP.) It is seen that the NNP describes energies of the small-size clusters well except for the smallest 4-atom cluster. This indicates that the uncoordinated Si atoms are also sampled within the G-metaD trajectories with the high bias. However, the present NNP may not provide enough accuracy to delineate the energy ordering among various geometries of the same-number nanoclusters.

DISCUSSION

The showcase examples on GeTe and Si in the above demonstrate that the G-metaD can produce training sets that are comparable to those that experts collected elaborately. To compare the size of the dataset for MLP development, 256,000 (G-metaD), 347,711 (ref. [8]), and around 1,000,000 (ref. [9]) atomic environments are sampled in GeTe, and 192,000 (G-metaD) and 171,815 (ref. [17]) in Si. Thus, the numbers of training points are similar to the previous state-of-the-art MLPs. This confirms that G-metaD can generate diverse and relevant configurations semi-automatically, which will expedite the development of MLPs by mitigating technicalities of choosing reference structures. However, a limited number of G-metaD trajectories may not provide full accuracy for every region of PES as observed in the example of Si wherein the EOS of some polymorphs is inaccurate. Therefore, we advise that practitioners augment the training set if high accuracy is necessary for specific configurations. For example, by including additional G-metaD trajectories starting with the same diamond structure but under constant pressures of 10–20 GPa, we could significantly improve EOS of high-pressure phases such as hcp and bc8 phases.

One can also utilize the G-metaD in complementing the traditional sampling style: since MLPs are essentially interpolative algorithms, the prediction error increases rapidly with structures outside the training domain. The MD-based sampling rarely explores high-energy regions and so the trained MLP is vulnerable to failures in long-term, large-scale simulations because some atoms may visit untrained regions eventually. This can be partly resolved by a weighting scheme[52] but the present G-metaD can provide a more robust solution. That is to say, after preparing a training set based on the traditional approach, practitioners may augment the training set by adding G-metaD trajectories, which extra-samples high-energy regions relevant for the simulation. This will achieve both high accuracy and stability of subsequent simulations.

In some cases, it is useful to apply a partial-bias G-metaD in which only a few selected atoms contribute to the total bias potential. For instance, to sample various sites of an interstitial atom (self or dopant types), one can add the interstitial atom into the crystalline bulk and apply the G-metaD only to the interstitial atom. This will enhance sampling of defective structures embedded in the crystalline bulk, which would not be feasible if the biasing force drives all atoms out of the crystalline structure simultaneously. For instance, diffusion paths of Li within a solid would be sampled efficiently by the partial-bias G-metaD[53].

We comment on an interesting connection between G-metaD and GAP; they both measure distances in the descriptor space in the same way, the former to previously visited points, the latter to fitting dataset configurations. Therefore, the local atomic environment which is not included in the dataset (or previous trajectories) has low bias potentials in G-metaD and high prediction variance (uncertainty) in GAP[11,54].

About the computational cost, the G-metaD takes about three times longer than the corresponding MD in the case of 20-ps simulations of GeTe. The present implementation of G-metaD operates the client-server mode between LAMMPS and VASP, and the read-write time of wave-function files accounts for one third of the total computation time. This could be alleviated by implementing the G-metaD directly into the ab initio program. Another source of higher computational loads is the computation of bias forces in Eq. (2). Unlike typical metaD, the CVs in G-metaD have large dimensions of 50–100 and the bias potential in Eq. (3) is contributed by every atom in the system. As a result, the computational time of bias calculations becomes significant as the G-metaD proceeds. For instance, it accounts for 20% of the computation time on average at 0–10 ps, which increases to 43% at 10–20 ps. By reducing the dimension of the **G** vector used in G-metaD, it would be possible to increase the computational speed substantially.

METHODS

Density-functional theory calculations

The reference DFT calculations are performed with Vienna Ab initio Simulation Package (VASP)[35] using projector augmented-wave pseudo-potentials[55]. The generalized gradient approximation is used for the exchange-correlation energy of electrons[56]. In the case of GeTe, we include a parameterized van der Waals interaction[57,58]. The temperature is controlled by the Nosé-Hoover thermostat and a time step of 2 fs is used. In MD or G-metaD simulations for H:Pt(111), the cutoff energy of 350 eV and **k**-point grid of $5 \times 5 \times 1$ are used. The G-metaD simulations with GeTe and Si are carried out with the cutoff energy of 300 and 250 eV, respectively, and the **k**-point grids are varied with a spacing of 0.4 Å^{-1} to maintain the computational consistency during the large volume change.

In obtaining reference energies, forces, and stress tensors of sampled structures, we perform one-shot DFT calculations with tighter parameters such that the total energy and atomic forces converge within 1.5 meV/atom and 0.04 eV/Å, respectively, for randomly sampled G-metaD snapshots. The resulting cutoff energy and **k**-point spacing are 400 eV and 0.3 Å^{-1} for GeTe, and 350 eV and 0.157 Å^{-1} for Si, respectively.

Neural network potential

The NNPs are trained by using SIMPLE-NN[34]. In training NNPs, the reference DFT data are split randomly into the training and validation sets with 9:1 ratio for H:Pt(111) and 19:1 ratio for GeTe and Si. We use the atom-centered symmetry function vector (**G**) to represent local environments[27]. The symmetry function vector **G** consists of radial (G_2) and angular components (G_4 and G_5) with cutoff radii of 3.5−8.0 Å. For training GeTe and Si, symmetry-function parameters are selected from a large pool of 233 sets by using the CUR method[59,60]. In the process of CUR, each symmetry function is penalized by rough estimates of evaluation costs based on the cutoff radius and the function type (radial or angular), thereby selecting the most cost-effective set of parameters. The CUR selection is terminated when the error (ε), defined in the following, drops below a certain threshold (0.001 and 0.002 for GeTe and Si, respectively):

$$\varepsilon = \left\| \mathbf{A} - \widetilde{\mathbf{A}} \right\|_F / \|\mathbf{A}\|_F, \tag{5}$$

where **A** is the original feature matrix constructed from 233 parameters, and **Ã** is the reduced feature matrix from selected parameters, and $\|\mathbf{A}\|_F$ is the Frobenius norm of matrix **A**. As a result, 61, 103, and 47 symmetry functions are selected for Ge, Te, and Si, respectively. For H:Pt(111), 70 parameters with a constant cutoff of 6.0 Å are selected without applying the CUR method.

For the NNP architecture, we adopt atomic neural networks with two hidden layers. The number of nodes per hidden layer is optimized with respect to the RMSE of the training set. As a result, each hidden layer consists of 30 nodes for H:Pt(111) and Si, and 60 nodes for GeTe. Since decorrelating the input vector benefits training quality and convergence speed, we transform the input vector by principal component analysis without dimensional reduction. After the transformation, variances of the vector components are normalized by whitening.

The training is performed with momentum-based Adam optimizer[61] with minibatch (the batch size of 20), which balances between

performance and computational costs. The initial learning rate is 0.0001 and reduced exponentially. The loss function (Γ) is formulated as follows:

$$\Gamma = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{E_i^{DFT}-E_i^{NNP}}{N_i}\right)^2 + \frac{\mu_1}{3}\frac{1}{\sum_{i=1}^{M}N_i}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left|\mathbf{F}_{ij}^{DFT}-\mathbf{F}_{ij}^{NNP}\right|^2 \tag{6}$$

$$+ \frac{\mu_2}{3}\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{3}\left|S_{ik}^{DFT}-S_{ik}^{NNP}\right|^2 + \frac{\mu_3}{3}\frac{1}{M}\sum_{i=1}^{M}\sum_{k=4}^{6}\left|S_{ik}^{DFT}-S_{ik}^{NNP}\right|^2,$$

where M is the total number of structures in the training set, N_i is the number of atoms in the ith structure. In Eq. (6), $E_i^{DFT(NNP)}$, $\mathbf{F}_{ij}^{DFT(NNP)}$, and $S_{ik}^{DFT(NNP)}$ for the ith structure indicate the total energy, atomic force of the jth atom, and the kth component ($k = 1$–6) of the virial stress, respectively. The scaling parameters μ_1, μ_2, and μ_3 in Eq. (6) control the weights of the force, normal ($k = 1$–3), and shear ($k = 4$–6) stress terms relative to the energy term, respectively. Since the shear components are usually smaller than normal ones, employing different scaling parameters μ_2 and μ_3 improve accuracy in the shear modulus. The training and validation errors are similar and their RMSE values for each system are noted in the main text.

Mahalanobis distance

We utilize the Mahalanobis distance to classify $\mathbf{G}$ vectors in GeTe structures into a point group θ (crystal, amorphous, or liquid phase)[46]. The Mahalanobis distance (d) measures the distance between a certain data point $\mathbf{x}$ and the center of the data point group θ in a multidimensional space[45]. It is calculated as:

$$d(\mathbf{x}, \theta) = \sqrt{(\mathbf{x}-\mu_\theta)^T\sum_\theta^{-1}(\mathbf{x}-\mu_\theta)}, \tag{7}$$

where μ_θ and Σ_θ are the mean and covariance matrix of the point group θ, respectively. When all axes are independent, Σ_θ becomes an identity matrix and the Mahalanobis distance is equal to the Euclidean one. In order to classify the unlabeled $\mathbf{G}$ vector, we first prepare reference data points for the GeTe phases from MD simulations: rock-salt crystal at 700 K, amorphous at 500 K (2 structures), and liquid at 1000 K. Then we randomly select 4,000 points from each MD simulation and label them as corresponding groups. Then, a $\mathbf{G}$ vector is classified into one of three phases (θ^*) for which it has the shortest d:

$$\theta^* = \underset{\theta \in \{C, A, L\}}{\arg\min} d(\mathbf{x}, \theta), \tag{8}$$

where C, A, and L indicate crystal, amorphous, and liquid structures, respectively. In addition, if d of a $\mathbf{G}$ vector is longer than that of the outermost $\mathbf{G}$ vector in each phase, it remains 'unclassified'. That is to say, if $\mathbf{x}^*$ is an unclassified $\mathbf{G}$ vector, the following relation holds:

$$\forall \theta, d(\mathbf{x}^*, \theta) > \max_{\mathbf{x} \in \theta}\{d(\mathbf{x}, \theta)\}. \tag{9}$$

Thus, surface structures or unary phases are unclassified. We also check the consistency by using d as a phase classifier: d can map $\mathbf{G}$ vectors from crystal, amorphous, and liquid into its original label with the accuracy of 92.2%, 78.8% and 100%, respectively. Most of the mislabeled $\mathbf{G}$ vectors are mapped into the liquid phase due to its largest variance.

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary information files.

CODE AVAILABILITY

The code for G-metaD is available on https://github.com/MDIL-SNU/G-metaD.

REFERENCES

1. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
2. Yoo, D. et al. Atomic energy mapping of neural network potential. *Phys. Rev. Mater.* **3**, 093802 (2019).
3. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
4. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
5. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model Simul.* **14**, 1153–1173 (2016).
6. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
7. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
8. Lee, D., Lee, K., Yoo, D., Jeong, W. & Han, S. Crystallization of amorphous GeTe simulated by neural network potential addressing medium-range order. *Comput. Mater. Sci.* **181**, 109725 (2020).
9. Sosso, G. C. et al. Fast crystallization of the phase change compound GeTe by large-scale molecular dynamics simulations. *J. Phys. Chem. Lett.* **4**, 4241–4246 (2013).
10. Mocanu, F. C. et al. Modeling the phase-change memory material, $Ge_2Sb_2Te_5$, with a Machine-learned interatomic potential. *J. Phys. Chem. B* **122**, 8998–9006 (2018).
11. Jeong, W., Yoo, D., Lee, K., Jung, J. & Han, S. Efficient atomic-resolution uncertainty estimation for neural network potentials using a replica ensemble. *J. Phys. Chem. Lett.* **11**, 6090–6096 (2020).
12. Hellström, M., Quaranta, V. & Behler, J. One-dimensional vs. two-dimensional proton transport processes at solid–liquid zinc-oxide–water interfaces. *Chem. Sci.* **10**, 1232–1243 (2019).
13. Sun, G. & Sautet, P. Metastable structures in cluster catalysis from first-principles: structural ensemble in reaction conditions and metastability triggered reactivity. *J. Am. Chem. Soc.* **140**, 2812–2820 (2018).
14. Hong, C. et al. Training machine-learning potentials for crystal structure prediction using disordered structures. *Phys. Rev. B* **102**, 224104 (2020).
15. Ulissi, Z. W. et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. *ACS Catal.* **7**, 6600–6608 (2017).
16. Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G. & Michaelides, A. An accurate and transferable machine learning potential for carbon. *J. Chem. Phys.* **153**, 034702 (2020).
17. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
18. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
19. Zhang, L., Lin, D. Y., Wang, H., Car, R. & Weinan, E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
20. Sivaraman, G. et al. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Comput. Mater.* **6**, 1–8 (2020).
21. Tong, Q., Xue, L., Lv, J., Wang, Y. & Ma, Y. Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discuss* **211**, 31–43 (2018).
22. Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).
23. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **5**, 1–9 (2019).
24. Karabin, M. & Perez, D. An entropy-maximization approach to automated training set generation for interatomic potentials. *J. Chem. Phys.* **153**, 094110 (2020).
25. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562–12566 (2002).
26. Nishihara, Y., Hayashi, S. & Kato, S. A search for ligand diffusion pathway in myoglobin using a metadynamics simulation. *Chem. Phys. Lett.* **464**, 220–225 (2008).
27. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
28. Bonati, L. & Parrinello, M. Silicon liquid structure and crystal nucleation from ab initio deep metadynamics. *Phys. Rev. Lett.* **121**, 265701 (2018).
29. Herr, J. E., Yao, K., McIntyre, R., Toth, D. W. & Parkhill, J. Metadynamics for training neural network model chemistries: a competitive assessment. *J. Chem. Phys.* **148**, 241710 (2018).
30. Sosso, G. C., Miceli, G., Caravati, S., Behler, J. & Bernasconi, M. Neural network interatomic potential for the phase change material GeTe. *Phys. Rev. B* **85**, 174103 (2012).

31. Gabardi, S., Sosso, G. G., Behler, J. & Bernasconi, M. Priming effects in the crystallization of the phase change compound GeTe from atomistic simulations. *Faraday Discuss* **213**, 287–301 (2019).

32. Branduardi, D., Bussi, G. & Parrinello, M. Metadynamics with adaptive gaussians. *J. Chem. Theory Comput.* **8**, 2247–2254 (2012).

33. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).

34. Lee, K., Yoo, D., Jeong, W. & Han, S. SIMPLE-NN: an efficient package for training and executing neural-network interatomic potentials. *Comput. Phys. Commun.* **242**, 95–103 (2019).

35. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).

36. Li, C. & Baek, J.-B. Recent advances in noble metal (Pt, Ru, and Ir)-based electrocatalysts for efficient hydrogen evolution reaction. *ACS Omega* **5**, 31–40 (2020).

37. Taguchi, S., Aramata, A. & Enyo, M. Reduced CO2 on polycrystalline Pd and Pt electrodes in neutral solution: electrochemical and in situ Fourier transform IR studies. *J. Electroanal. Chem.* **372**, 161–169 (1994).

38. Wilde, M. et al. Influence of carbon deposition on the hydrogen distribution in Pd nanoparticles and their reactivity in olefin hydrogenation. *Angew. Chem. Int. Ed. Engl.* **47**, 9289–9293 (2008).

39. Zhai, F., Li, Y., Yang, Y., Jiang, S. & Shen, X. Abnormal subsurface hydrogen diffusion behaviors in heterogeneous hydrogenation reactions. *J. Chem. Phys.* **149**, 174704 (2018).

40. Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).

41. Smidstrup, S., Pedersen, A., Stokbro, K. & Jónsson, H. Improved initial guess for minimum energy path calculations. *J. Chem. Phys.* **140**, 214106 (2014).

42. Bădescu, S. C. et al. Energetics and vibrational states for hydrogen on Pt(111). *Phys. Rev. Lett.* **88**, 136101 (2002).

43. Ferrin, P., Kandoi, S., Nilekar, A. U. & Mavrikakis, M. Hydrogen adsorption, absorption and diffusion on and in transition metal surfaces: A DFT study. *Surf. Sci.* **606**, 679–689 (2012).

44. Raoux, S., Wełnic, W. & Ielmini, D. Phase change materials and their application to nonvolatile memories. *Chem. Rev.* **110**, 240–267 (2010).

45. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition* (John Wiley & Sons, Hoboken, NJ, 2004.)

46. Maesschalck, R. D., De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **50**, 1–18 (2000).

47. Roux, S. L., Le Roux, S. & Jund, P. Ring statistics analysis of topological networks: new approach and application to amorphous GeS_2 and SiO_2 systems. *Comput. Mater. Sci.* **49**, 70–83 (2010).

48. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28**, 784 (1983).

49. Rohlfing, C. M. & Raghavachari, K. A theoretical study of small silicon clusters using an effective core potential. *Chem. Phys. Lett.* **167**, 559–565 (1990).

50. Ho, K.-M. et al. Structures of medium-sized silicon clusters. *Nature* **392**, 582–585 (1998).

51. Tomanek, D. & Schluter, M. A. Calculation of magic numbers and the stability of small Si clusters. *Phys. Rev. Lett.* **56**, 1055–1058 (1986).

52. Jeong, W., Lee, K., Yoo, D., Lee, D. & Han, S. Toward reliable and transferable machine learning potentials: uniform training by overcoming sampling bias. *J. Phys. Chem. C* **122**, 22790–22795 (2018).

53. Wang, C., Aoyagi, K., Wisesa, P. & Mueller, T. Lithium ion conduction in cathode coating materials from on-the-fly machine learning. *Chem. Mater.* **32**, 3741–3752 (2020).

54. Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 1–11 (2020).

55. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

56. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

57. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).

58. Sosso, G. C., Behler, J. & Bernasconi, M. Breakdown of Stokes-Einstein relation in the supercooled liquid state of phase change materials. *Phys. Status Solidi B* **249**, 1880–1885 (2012).

59. Imbalzano, G. et al. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).

60. Mahoney, M. W. & Drineas, P. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 697–702 (2009).

61. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv*. Preprint at https://arxiv.org/abs/1412.6980 (2014).

62. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the open visualization tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2010).

63. Chadi, D. J. Reexamination of the Si(100) surface reconstruction. *Appl. Opt.* **19**, 3971 (1980).

64. Menon, M., Lathiotakis, N. N. & Andriotis, A. N. The reconstruction of the Si(110) surface and its interaction with Si adatoms. *Phys. Rev. B.* **56**, 1412 (1997).

65. Solares, S. D. et al. Density functional theory study of the geometry, energetics, and reconstruction process of Si(111) surfaces. *Langmuir* **21**, 12404–12414 (2005).

AUTHOR CONTRIBUTIONS

D.Y. put forward the original idea of G-metaD and further refined the method with J.J., W.J., and S.H. D.Y. and J.J. contributed equally to this paper. S.H. organized the whole project. All the authors participated in writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00595-5.

Correspondence and requests for materials should be addressed to S.H.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.