Active Δ-learning with universal potentials for global structure optimization

Joe Pitfield,[*] Mads-Peter Verner Christiansen, and Bjørk Hammer[†]
Center for Interstellar Catalysis, Department of Physics and Astronomy,
Aarhus University, DK-8000 Aarhus C, Denmak

(Dated: July 25, 2025)

Universal machine learning interatomic potentials (uMLIPs) have recently been formulated and shown to generalize well. When applied out-of-sample, further data collection for improvement of the uMLIPs may, however, be required. In this work we demonstrate that, whenever the envisaged use of the MLIPs is global optimization, the data acquisition can follow an active learning scheme in which a gradually updated uMLIP directs the finding of new structures, which are subsequently evaluated at the density functional theory (DFT) level. In the scheme, we augment foundation models using a Δ-model based on this new data using local SOAP-descriptors, Gaussian kernels, and a sparse Gaussian Process Regression model. We compare the efficacy of the approach with different global optimization algorithms, Random Structure Search, Basin Hopping, a Bayesian approach with competitive candidates (GOFEE), and a replica exchange formulation (REX). We further compare several foundation models, CHGNet, MACE-MP0, and MACE-MPA. The test systems are silver-sulfur clusters and sulfur-induced surface reconstructions on Ag(111) and Ag(100). Judged by the fidelity of identifying global minima, active learning with GPR-based Δ-models appears to be a robust approach. Judged by the total CPU time spent, the REX approach stands out as being the most efficient.

I. INTRODUCTION

Over the past two decades, the fields of molecular, colloid, and materials science have supported the development of highly flexible machine learning interatomic potentials (MLIPs) [1–4]. Formulated using frameworks such as feed-forward deep neural networks [5] or Gaussian Process Regression models [6], such MLIPs have proven highly efficient for atomistic simulations allowing, for e.g. larger and more complex systems [7] and longer timescales compared to simulations performed fully with density functional theory (DFT) [8–11]. The data used to train such models is collected according to various protocols [12], including random sampling [13, 14], global optimization searches [15, 16], normal mode analysis [17], molecular dynamics-driven sampling [18, 19], and saddle point searching [20]. Active learning schemes have been invoked, in which the models are retrained upon assembly of data for which the models signal uncertainty [21–31].

Recently, equivariant graph-neural network models have been introduced for MLIPs [32, 33]. These models encode the vectorial properties of the local environment and embed information about farther environments via message passing. Such models are generally more data efficient [34] and have supported the introduction of foundation models [35–37], where general purpose, "universal" MLIPs (uMLIPs) are trained on huge preassembled databases spanning the entire Periodic Table and including compounds bonded covalently, ionically, dispersively, and as metals. A swathe of recent advancements in the field of foundation models have shown rapid improvement in the benchmarks of these models and their suitability for direct application [38, 39].

A common procedure for making such advancements and improving foundation MLIPs and MLIPs in general, focuses on adding more diverse training data [40]. An example of this is the emergence of the successive models MACE-MP0 and MACE-MPA [37]. The training datasets are for both models that of the Materials Project database (MPtrj) [36], and in the case of MACE-MPA a portion of the Alexandria database [41]. The larger training dataset for MACE-MPA causes the model to perform more accurately in predictions across the Matbench Discovery benchmark [42]. An alternative approach for improving foundation MLIPs is via loss function engineering and change of design philosophy. The FAIR META eSEN-30M-OAM model [43] shows iterative improvement by acknowledging the relevance of quality metrics, such as KSRME (symmetric relative mean error in the phonon contribution to the thermal conductivity, which correlates with the smoothness of the PES) in describing when a potential is truly performative.

Whilst it does not fall within the scope of most materials scientists to perform such bespoke model training from scratch (more so than ever with models such as the 1.6 billion parameter UMA model [44]), other methods of either changing or adding to existing models to fit a given application, do. When changing an existing model is limited in scope to adding more data (and potentially changing the loss function), the approach is commonly referred to as "fine-tuning".

Fine-tuning has shown to be effective in improving the efficacy of models, so much so that the training methodology for some of the best performing models (eSEN) actually included fine-tuning explicitly. It has been found that such fine-tuning is effective at remedying systematic problems such as softening of the PES [40]. Other systematic examples, such as surface energies [45], alloy mixing [46], thermal conductivity [47], phonons [48] and

[*] joepitfield@gmail.com
[†] hammer@phys.au.dk

sublimation enthalpies [49], have shown similar improvement under fine-tuning with 10s - 1000s of supplementary datapoints, depending on the property.

Crucially, the current metrics for fine-tuning or training large models in general do not place any emphasis on the energetic ordering of low energy structures. Often, the global minimum energy configuration for a system exists outside of the intuitive, thanks to effects which one can describe as phenomenological rather than systematic. uMLIPs, without proportional incentive to understand such phenomena, can become deficient in the discovery regime. Studying silicate clusters and ultra-thin surface-oxides on Ag(111), we have previously demonstrated that fine-tuning a uMLIP is reliably able to correct relatively minor structural inconsistencies and yield correct ordering of low-energy configurations [50]. Notably, this required a dataset obtained through active learning enhanced exploration of the PES. This procedure would have been difficult and expensive to carry out with iterative fine-tuning.

An alternative means to correct a uMLIP is that of an additive Δ-model [50], which is agnostic to several elements of the original uMLIP, including i) the choice of loss function, ii) the network architecture, and iii) the data used in the original training. Such a correction is often able to encode phenomenological effects, which are not well described by the uMLIP. More so, where such effects could cause catastrophic forgetting and deteriorate the performance of the model, were they to be provided as small amounts of training data for fine-tuning, a Δ-model is less prone to influence the interpolation domain. A further benefit of a scheme with a Δ-model is, that the training is computationally cheap and can be applied often, not requiring the collection of large batches of new data before updating the potential.

In the current work, we investigate the viability of establishing a reliable Δ-model corrected uMLIP via an active-learning scheme. The use-case envisaged for the resulting corrected uMLIP is global structure optimization, and hence the data-collection is guided by structures realized during such optimizations. Different algorithms are considered, Random Structure Search, Basin Hopping, and some more elaborate ones introduced below. With these methods, the global minimum DFT energy structures (GMs) of a range of $[Ag_2S]_X$ clusters can be found via active-learning of a Δ-model added to a uMLIP. When using different uMLIPs for active learning – CHGNet, MACE-MP0, and MACE-MPA – we find little dependence on the rate of identifying the GMs for the clusters, while for sulfur-induced reconstructions of silver surfaces some differences are found, with MACE-MPA leading to the fastest discovery of the global minimum energy structure. We will discuss this observation in terms of MACE-MPA having encoded important Ag-S motifs more accurately, and hence needing less data in the active-learning searches.

The paper is outlined as follows: In the methodology section we introduce the three elements of the active learning: i) the different uMLIPs used, ii) the Gaussian Process Regression-based Δ-model, and iii) and the four different global optimization algorithms. The results section starts with a discussion of prior understanding of the uMLIPs without any correction. It proceeds by presenting how the active learning performs for $[Ag_2S]_X$ clusters when varying the optimization method. Then the use of different uMLIPs in the context of active learning is considered, first for the clusters and next for the sulfur induced surface reconstructions. As the final topic in the results section, the use of prior collected data for pre-correction of the uMLIPs is considered. The paper ends with a discussion and details on data and code availability.

II. METHODOLOGY

A. Universal MLIPs and training datasets

MPTrj consists primarily of bulk structural information, namely 1.58 million unique structures and relaxation trajectories, computed at the PBE [51] level of DFT. The Alexandria dataset is more diverse, containing 1D and 2D periodic systems alongside 3D bulk structures, providing more chemical insight into finite size effects particularly relevant to those investigated in this work. Overall, the dataset contains 30+ million structures, although this dataset is often sub-sampled to avoid structures being overrepresented (the sub-sampled dataset is *only* 10 million structures). These structures are computed with the PBEsol [52] and SCAN [53] XC functionals.

In this work, we will examine three different uMLIPs namely CHGNet [36], MACE-MP0-large (henceforth referred to simply as MACE-MP0), and MACE-MPA-0 (MACE-MPA). CHGNet is a graph neural network potential, in which site based magnetic moments have been incorporated into the training data. The architecture of CHGNet involves both an atom graph where nodes are atoms carrying atomic embeddings and an auxiliary bond graph where the nodes are bonds and edges carry angular information. Using interactions between these elements CHGNet incorporates angular information into the atomic embeddings. Ultimately, the atomic embeddings are used to predict total energies and magnetic moments in addition to forces and stresses through the use of automatic differentiation of the total energy.

MACE is also a graph neural network, but unlike CHGNet it is equivariant with higher bond-order information gathered through tensor products involving directional information decomposed through spherical harmonics - making it capable of directly encoding angular information, dihedrals and beyond. These architectural differences have proven to improve performance across various benchmarks as evidenced by MatBench discovery [42]. We use two variants of MACE, namely MACE-MP0 trained solely on the MPTrj dataset that CHGNet is also trained on, and MACE-MPA which is trained on MPTrj and a subset of the Alexandria dataset. The larger training set used for the MACE-MPA uMLIP increases it performances on the aforementioned benchmarks compared to MACE-MP0. We employ the 'large' version of MACE-MP0, with $\sim$ 16 million parameters and the medium version of MACE-MPA, with $\sim$ 9 million parameters. Both models use the same building blocks and we attribute the majority of the differences in their behaviour that we observe to the difference in training set while acknowledging that we cannot prove this to be the case. In particular we later investigate the energetic ordering of clusters and observe significant differences in the ordering between MACE-MP0 and MACE-MPA, which we contribute to the training set differences but could also be partially caused by the architectural differences or even the random initialization of the network parameters.

B. Δ-model

Previously, Δ-models have been used in bridging the gap between levels of theory [54], including excited states [55] and prediction of correlation energies [56]. Their potential application in translating between levels of prediction has been demonstrated to be effective [57]. Here, we apply that same translation between uMLIP predictions and DFT predictions using a Δ-model [50]. The corrected energy prediction for a structure, $\mathcal{M}$ (specifying super cell, atomic identities, and positions), is given by:

$$E_{\rm model}(\mathcal{M}) = E_{\rm uMLIP}(\mathcal{M}) + E_{\Delta}(\mathcal{M}) \tag{1}$$

where $E_{uMLIP}(\mathcal{M})$ is the unmodified uMLIP, and where $E_{\Delta}(\mathcal{M})$ is the correction fitted to the residuals between the data and the incorrect uMLIP predictions.

In this work we evaluate $E_{\Delta}(\mathcal{M})$ as a sparse Gaussian Process Regression based on a Gaussian kernel evaluating the similarities between the local atomic SOAP [58] descriptors of the query structure and the training structures for the Δ-model. See Refs. 15 and 59 for details on the GPR model.

Since the Δ-correction in the present work is described by a Gaussian Process (GPR), an interesting analogy exists, namely that Eq. 1 is equivalent to a GPR model in which the uMLIP serves as a prior [50, 60]. The analogy, however, ceases to hold in cases where the uncertainty of the uMLIP can be quantified and included in the GP, or when the Δ-correction is modeled with a neural network [61].

The training structures are comprised of the structures emerging from the global optimization algorithms detailed in the next subsection. In cases where data is available prior to the searches, some pretraining data will be included in the pool of training data for the sparse-GPR model, while being excluded from the pool of structures undergoing iterative improvement according to the global optimization scheme, where relevant.

The SOAP descriptor is implemented in DScribe [62, 63]. We use $(n, l) = 3, 2$ and $r_{cut} = 7$ Å. The sparse-GPR [15] is implemented in AGOX [64]. We use 1000 inducing points and train only on energies to keep the size of the kernel matrix manageable.

C. Global optimization methods

Many choices exist for which global optimization method to employ for the active learned data collection of the Δ-model introduced above. We have chosen a collection of methods ranging from the almost completely unbiased random structure search, to a Monte Carlo based Basin Hopping method, on towards more elaborate Bayesian- and parallel tempering-inspired methods. This allows for an assessment of the dependence of the active learning scheme on the optimization method. All methods introduced will use the Δ-corrected uMLIP model for surrogate relaxation and perform the model update whenever new DFT data becomes available. Relaxations are performed for up to a certain number of steps, or until a force convergence threshold is reached. For RSS and Basing Hopping, up to 500 relaxation steps are performed. For the Bayesian method we use up to 100 relaxation steps, and for the parallel tempering-inspired method, 30. The force convergence threshold in almost all cases is 0.05 eV/Å, with the exception of the Ag(100)-($\sqrt{17} \times \sqrt{17}$), for which a more stringent 0.025 eV/Å is used. Relaxations are conducted with the BFGS [65] local optimization technique implemented in ASE [66]. Figure 1 presents a graphical overview of the global optimization methods used. All of the optimization methods we employ are implemented in the modular framework of the AGOX global optimization package [64].

Common to all methods is that atoms are confined to *confinement cells* both when they are introduced and during relaxation. For each of the $[Ag_2S]_X$ structures, a $20 \times 20 \times 20$ Å unit cell houses a $15 \times 15 \times 15$ Å confinement cell. The sulfur-induced surface reconstruction searches are done over 4 layer slabs of static Ag atoms in either Ag(111)-($\sqrt{7} \times \sqrt{7}$) or Ag(100)-($\sqrt{17} \times \sqrt{17}$) cells using (6×6) and (2×2) **k**-point grids for sampling of the corresponding 2D Brillouin zones. The reconstructions are built from 3×Ag+3×S and 12×Ag+8×S atoms introduced in confinement cells with the same dimensionality as the lattices in the xy directions and having a z component of 4 Å. All DFT calculations are performed with plane wave GPAW [67] (500 eV cutoff) and employing the PBE XC-functional [51].

1. *MLIP-assisted Random Structure Search (RSS)*

Random structure search is one of, if not *the*, simplest structure search method widely applied in the history of

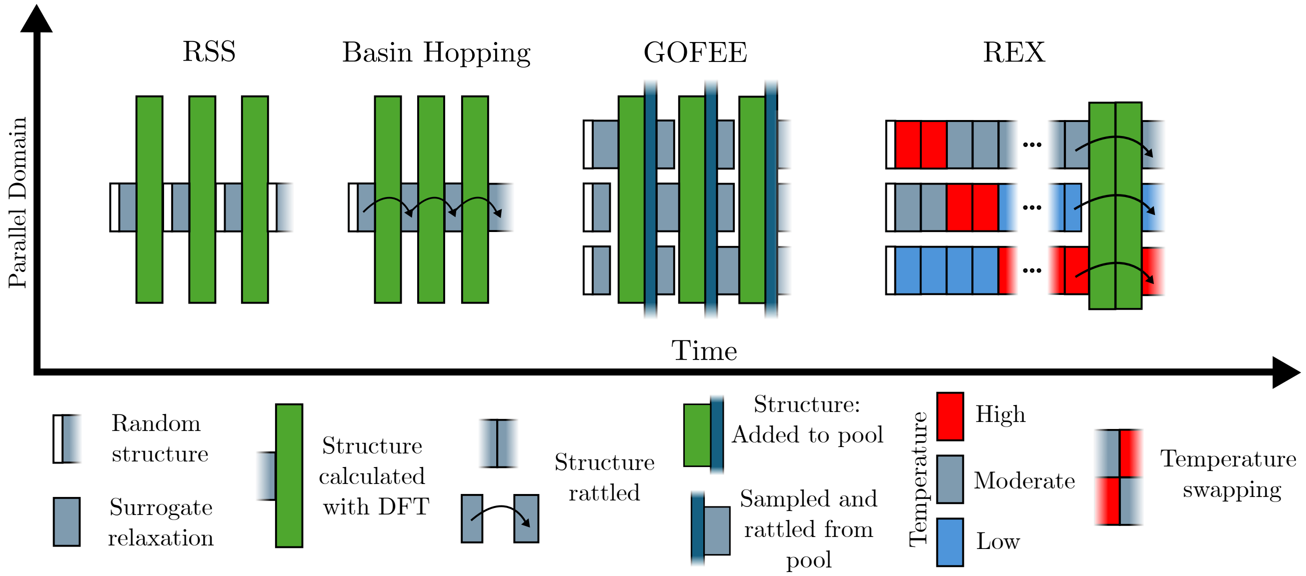

FIG. 1. Schematic outlining the data collection and progression schemes of RSS, Basin Hopping, GOFEE, and REX. The key highlights the processes represented in the schematic. The conditioned acceptances, Eqs. 2 and 4, of rattled and surrogate relaxed structures in Basin Hopping and REX are left out for clarity. The swapping events of REX are shown as walkers interchanging temperatures. In AGOX, this is coded as structures swapping structures.

computational materials science[68]. Atoms are placed within a cell at random whilst still accounting for reasonable bond lengths. Some small constraints are placed on the definition of random (no two atoms can be placed too close together, atoms must be placed within some distance of other atoms). In the present work, we require that no two atoms be placed closer together than 60% of the average of the covalent radius of the two species, and that each atom must have at least one atom within 3 times this distance.

This structure, after surrogate relaxation, is then evaluated with first principles methods. As seen in Fig 1, each instance of RSS is entirely independent from others, with the exception that the surrogate potential landscape is built from predecessor structures, introducing slight correlation between structures over time.

2. *MLIP-assisted Basin Hopping (BH)*

Basin Hopping is a Monte Carlo based technique in which structural walkers are evolved via random perturbations and relaxations [69]. In our implementation, all atoms are rattled according to a uniform random distribution within a sphere around their original position (whilst still accounting for reasonable bond lengths). The radius of the sphere is determined by the displacement magnitude, which for Basin Hopping is set to a static 3 Å. Structures are subsequently relaxed in the Δ-corrected uMLIP. The new candidate, $\mathcal{M}_{\text{new}}$ is accepted according to the Metropolis Monte Carlo criterion probability, given by:

$$P_{\text{accept}}(\mathcal{M}_{\text{new}}|\mathcal{M}) = \min\left(1, \exp\left[-\frac{E_{\text{DFT}}(\mathcal{M}_{\text{new}}) - E_{\text{DFT}}(\mathcal{M})}{k_B T}\right]\right), \quad (2)$$

where $E_{\text{DFT}}(\mathcal{M})$ is the energy of the current structure, $E_{\text{DFT}}(\mathcal{M}_{\text{new}})$ is the energy of the proposed state, and the product of the Boltzmann constant and the temperature, $k_B T$, which is 0.15 eV. Basin Hopping introduces explicit correlation in the time domain between structures considered – i.e. structures depend on their structural predecessors, and are commonly referred to as walkers to promote this notion. These are illustrated by arrows on the Basin Hopping diagram shown in Fig 1. Regardless of the acceptance and propagation of the proposed structure through further iterations of the algorithm, its energy is calculated in DFT and added to the training set.

3. *GOFEE*

Global Optimization for First-principles Energy Expressions (GOFEE) introduces the concept that multiple walkers can be treated in parallel, as seen in Fig. 1. GOFEE leverages principles of Bayesian statistics when creating new structural candidates. It does so in two ways. Firstly, a set of 10 structures are drawn from the pool of DFT evaluated structures. They are selected according to a K-means clustering of all structures in the pool, where each cluster contributes its lowest energy structure to a walker. The walkers are rattled as in

Basin Hopping, but then relaxed in the lower confidence bound,

$$LCB(\mathcal{M}) = E_{model}(\mathcal{M}) - \kappa\sigma(\mathcal{M})$$

where $\sigma(\mathcal{M})$ is the predicted uncertainty on E_{model}. κ is a constant, typically 2, which we maintain in this work. The structure with the lowest value in the lower confidence bound is then evaluated with DFT. This structure is finally added to the pool of evaluated structures, and a single DFT relaxation step performed. The process is then iteratively repeated. The process is then iteratively repeated. More details can be found in Ref. 70. The uncertainty is evaluated employing an ensemble of GPR models trained on data with artificial noise ($\sigma_p = 0.001$ eV/atom, $\sigma_l = 0.025$ eV/atom) added as proposed in Ref. 59.

4. ML-assisted Replica Exchange X (REX)

Like Basin Hopping, REX involves walkers which inherit the structure from the previous iteration, incrementally improving and exploring. Like GOFEE, it utilizes multiple instances of walkers progressing in parallel, sharing a surrogate potential between the walkers. This method also draws inspiration from replica exchange (RE) methodologies in materials science [71–76], particularly parallel tempering (PT) [77, 78]. Parallel tempering is a special case of RE wherein the temperature is the only differing parameter between coupled replicas, and is often applied with either Monte Carlo (MC) or molecular dynamics (MD) evolution, giving rise to the acronyms REMC and REMD, respectively. These methods are both formulated to provide thermal samples for a given potential and overcome local minima and obtain ergodic behavior by maintaining an ensemble of walkers evolved in parallel at different temperatures. The i'th and j'th walkers are subject to swapping events with probability:

$$P_{\mathrm{swap}}(i,j) = \min\Bigg(1,\, \exp\Big[(\frac{1}{k_BT_i} - \frac{1}{k_BT_j}) \cdot\big(E_{\mathrm{model}}(\mathcal{M}_j) - E_{\mathrm{model}}(\mathcal{M}_i)\big)\Big]\Bigg), \tag{3}$$

which can be implemented either by swapping the candidates between fixed-temperature walkers or by swapping the temperatures of the walkers that keep their structural candidates. After an accepted swapping event, the involved structures will propagate with altered temperatures. The rational is that in the potential energy landscape high-temperature walkers identify new regions of importance for sampling at lower temperature, and by making these regions known to low-temperature walkers, the low-temperature sampling converges faster. In a global optimization context, this means that the GM will be identified more efficiently.

In our implementation, we leverage the main features of RE, but use Basin Hopping walkers, i.e. walkers that are relaxed in the surrogate energy landscape of the Δ-corrected uMLIP before being subjected to the Metropolis Monte Carlo acceptance test, which is further based on the surrogate energies,

$$P_{\mathrm{accept}}(\mathcal{M}_{\mathrm{new}}|\mathcal{M}) = \min\Big(1,\, \exp\Big[-\frac{E_{\mathrm{model}}(\mathcal{M}_{\mathrm{new}}) - E_{\mathrm{model}}(\mathcal{M})}{k_BT}\Big]\Big). \tag{4}$$

Due to the relaxations involved, detailed balance will be violated, and hence rigorous thermal samples are not obtained by our RE implementation. However, the method still benefits fully from the exploratory potential of coupled walkers.

We dub our method Replica Exchange X (REX), where the X denotes the departure from previous methods, and the presence of other additions to the aforementioned algorithms, such as the uMLIP-based surrogate potential.

For each REX search, 10 walkers are instantiated with pseudo-random structures, exactly as with RSS. Each walker is assigned a value for k_BT according to a geometric series between 0.1 and 1. Similarly, each walker is assigned a different initial value for displacement magnitude, linearly between 0.1 and 5 Å. This determines the magnitude of attempted displacements, which are exactly as in Basin Hopping. The highest temperature walker always generates a pseudo-random structure. The displacement magnitude dynamically alters over the course of the calculation to target a 50% metropolis acceptance rate for new structures. 10 iterations of independent Basin Hopping events are performed between swapping attempts, and 30-50 between DFT calculations, where more complex systems are allowed more iterations between DFT calculations. When DFT is to be performed, 5 structures are randomly selected from the 9 lowest temperature walkers thereby avoiding the pseudo-random structure. This process is repeated until the allocated time budget has expired.

D. $[\mathrm{Ag_2S}]_X$ with uncorrected uMLIPs

III. RESULTS

The results of these methods when applied to a variety of systems is discussed here. We present these results by way of success curves [25, 64]. Each vertical increment of a success curve indicates that one independent search has identified the solution at the given x-coordinate (usually CPU time or number of search episodes). Many searches together then provide a statistical ensemble of success (with each individual search termed a repeat), which is more indifferent towards flukes. This can also be viewed as the integral of the histogram of success. For the nanoclusters, we calculate the spectral decomposition graph for the best DFT structure in each search [64]. Success

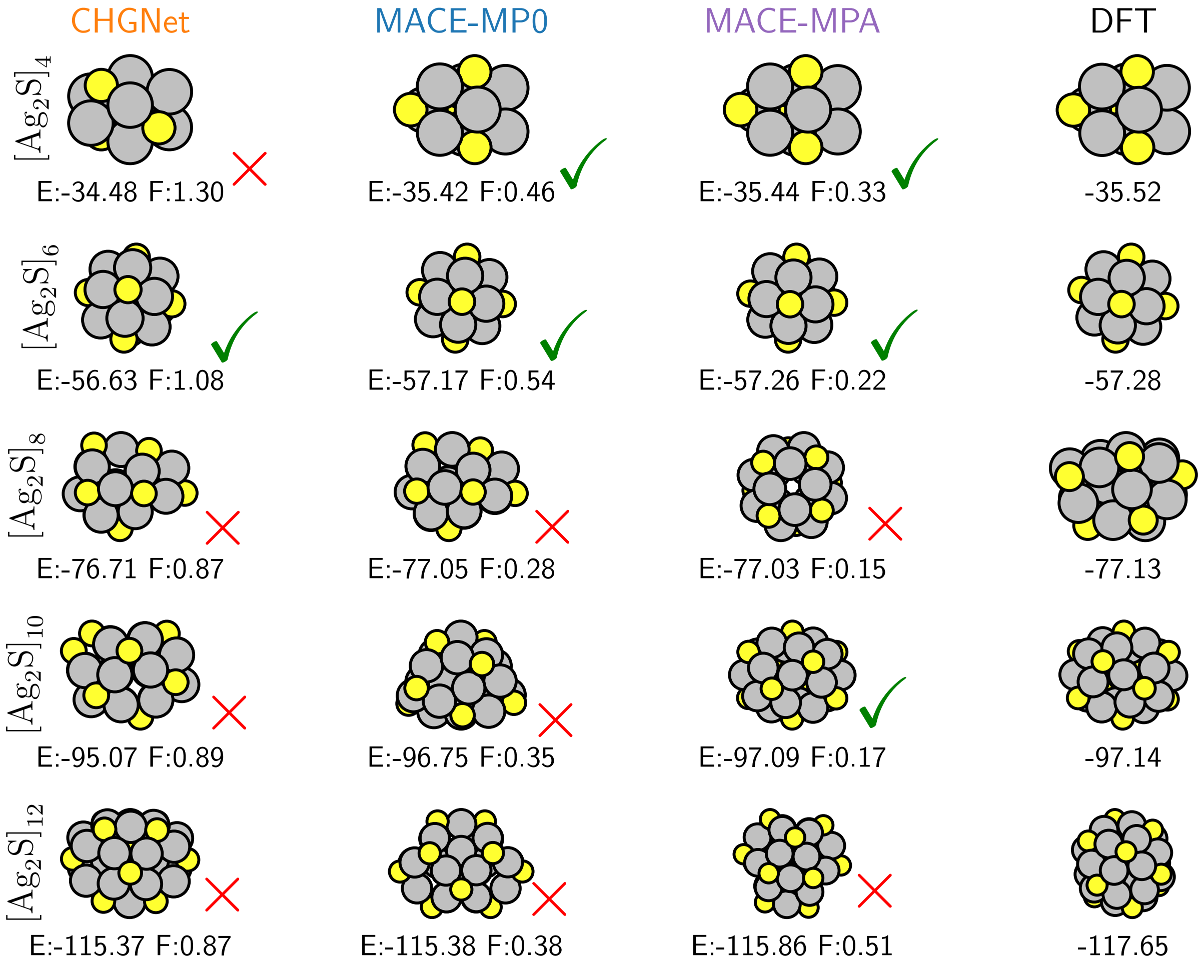

FIG. 2. Global minimum energy structures in CHGNet, MACE-MP0, MPA, and DFT. The energies (eV) and maximum atomic forces (eV/Å) calculated in DFT are shown underneath. Green ticks and red crosses indicate if the structure identified by the universal potential matches configurationally with the DFT GM structure of the corresponding stoichiometry.

is then defined as having a graph which is identical to that of the best structure. In the case of surface reconstructions, we define success as having an energy within a threshold of the solution. In order to obtain a comparable value for the CPU time, we ensure each search is performed on identical architecture and resources, namely 10 cores from 48 of 2 Intel Xeon Gold 6248R CPUs, such that the CPU time is equivalent to ten times the walltime. DFT calculations are run in parallel on all 10 CPUs. Surrogate calculations run on 1 CPU, and multiple will run in parallel across the 10 CPUs if allowed by the algorithm.

This section is organized as follows: We first discuss how well the three uMLIPs describe the low-energy conformers of $[Ag_2S]_X$ clusters for five values of X. In more than half of the combinations of uMLIP and cluster size, it is found that the global minimum uMLIP energy structure (uMLIP-GM) deviates significantly from the global minimum DFT energy structure (DFT-GM). The section proceeds to demonstrate that for a given uMLIP the DFT-GM can be found using the proposed active learning Δ-model. From this analysis, it is further established that as the cluster size and hence the complexity of the optimization problem increases, more advanced search algorithms must be used. The section moves on to consider the importance of the quality of the uMLIP. For the $[Ag_2S]_X$ clusters, the discrepancies between uMLIP-GMs and DFT-GMs do not appear to delay the finding of the DFT-GMs in our active learning scheme. This suggests that the problem of probing the proper configuration dominates the problem of correcting the uMLIP. For sulfur-induced surface reconstructions, the situation is reversed for the most complex problem, and MACE-MPA, which needs less correction to describe the DFT-GM also facilitates its faster finding. The section ends with a discussion of the usefulness of precorrecting a uMLIP using data from prior searches.

Figure 2 shows the uMLIP-GM predictions of the three universal potentials together with the DFT-GMs for $[Ag_2S]_X$ clusters having $X \in \{4, 6, 8, 10, 12\}$. The uMLIP-GMs were found by conducting exhaustive REX searches, without the active Δ-learning, while the DFT-GMs were compiled from the results of the active learning searches presented below in sections III A and III B. The cluster illustrations are annotated with the total DFT energies and the maximum magnitude of the atomic forces. A green tick or red cross indicate whether or not a DFT-based relaxation causes the structure to assume the DFT-GM configuration, respectively.

For the smaller clusters, $X \in \{4, 6\}$, the uMLIP-GMs and DFT-GMs agree except for the combination of CHGNet with $[Ag_2S]_4$. Interestingly, this wrong prediction involves a cluster of higher symmetry than the DFT-GM. In contrast, a view at the DFT-GM for $[Ag_2S]_6$ reveals that it is a highly symmetric structure. This coincides with all three uMLIPs predicting the correct structure for this cluster size, and hints that the network architectures and training datasets of the uMLIPs might favor high symmetry.

For the larger clusters, $X \in \{8, 10, 12\}$, only the MACE-MPA prediction for $[Ag_2S]_{10}$ is correct, while all other predictions are wrong. We associate the higher fail rate for the larger clusters with their configuration spaces being considerably larger, putting the uMLIPs to a more stringent test.

Focussing on the DFT-based forces evaluated for the uMLIP-GMs, there is a clear tendency of decreasing force magnitudes going from CHGNet, to MACE-MP0 and then MACE-MPA. Comparing CHGNet and MACE-MP0, the difference must originate from the network architecture, as they have been trained on the same materials project dataset. Comparing MACE-MP0 and MACE-MPA, that are based on more similar network architectures, the difference more likely lies in the datasets, and it is seen that adding the Alexandria dataset has the desired effect of leading to a more accurate uMLIP.

The DFT-GMs presented in Fig. 2 pertain to a DFT setting using the PBE XC functional and the uMLIPs are trained on PBE-based data, rendering the above comparison meaningful. In the literature, GM structures for other XC functionals have been reported. Using e.g. PBE0, a GM structure has been reported [79] which is similar to the MACE-MPA prediction for the $[Ag_2S]_8$ cluster shown in Fig. 2. Relaxing our MACE-MPA and DFT GM structures with PBE0 we confirm this result, but we also find that by including the D3 [80] van der Waals dispersion term and performing PBE0-D3 DFT calculations, the PBE DFT-GM again becomes the preferred structure. We stress, however, that since the currently used uMLIPs were trained on PBE data, the fair comparison is to other PBE based results. Hence, the green ticks in Fig. 2 indicate the uMLIP predicting the correct PBE DFT-GM structure. It is also noteworthy that the Δ-model strategy trivially extends to coupling uMLIPs trained on one level of theory to DFT calculations performed at another.

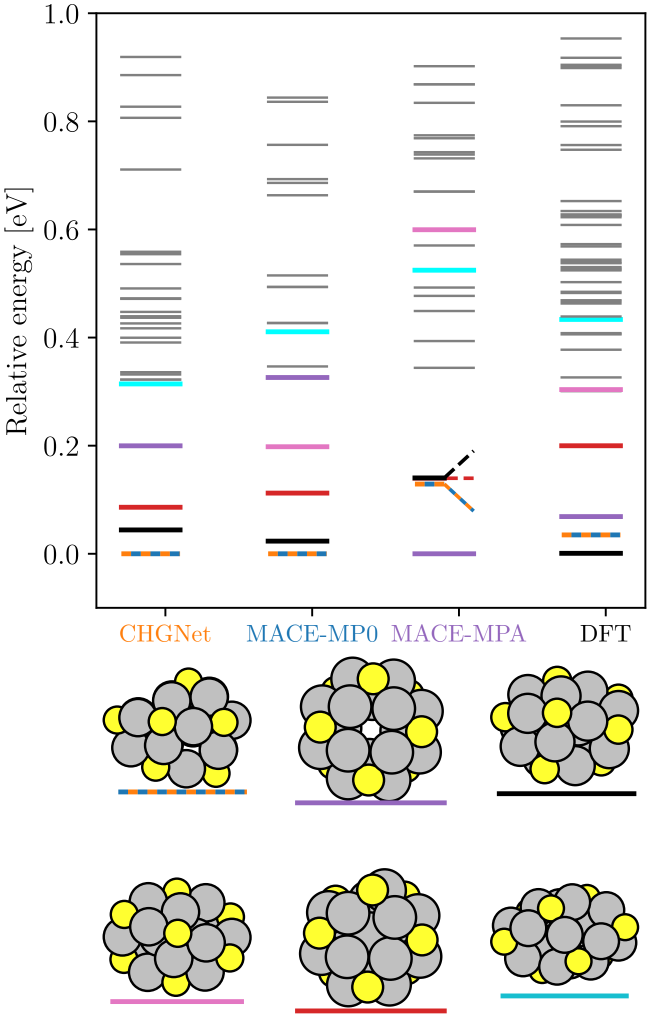

FIG. 3. Depicted are the relative energies of a subsampling of $[Ag_2S]_8$ structures relaxed in CHGNet, MACE-MP0, MACE-MPA and DFT. These energies are relative to the lowest energy obtained in any given potential. Structural diagrams are provided for a selection of structures according to color.

In order to provide a more comprehensive picture of quality of the uncorrected uMLIPs, Fig. 3 presents the relative stability of a set of different low-energy conformers of the $[Ag_2S]_8$ cluster. The structures are the uMLIP-GMs (one of which is shared between CHGNet and MACE-MP0), the DFT-GM, two other structurally distinct nanoclusters (red, pink, cyan) and a distribution of further structures (grey).

The structures in Fig. 3 were obtained from a set of REX searches either with active learning of a Δ-model corrected MACE-MP0 model to obtain local DFT minima, or without active learning in three sets of REX searches to identify local minima in each of the three uMLIPs considered. The resulting dataset is thus representative for both DFT and the individual potentials. Overall, 10500 structures were obtained in this way. We subsample this dataset (for which 80% of structures fall into one of four basins) to reduce repetition by selecting only 1 in 100 structures, resulting in 105 structures. Each structure is then relaxed in each uMLIP or in DFT and filtered for similarity to other structures. The resulting structures thus represent the relative ordering of minima as they appear in each potential.

Comparing the stability order of the various conformers within each uMLIP to that in DFT, it is seen that the relative order is highly sensitive to the model, as we demonstrate in Figure 2. Firstly, the CHGNet and MACE-MP0 structures are configurationally alike, residing in the same basin and differing only through finer structural parameters, and thus share a structural model and blue/orange coloring. The small differences between the MACE-MP0 and CHGNet structures are nonetheless significant enough to contribute to a large difference between the respective energy predictions. As seen in Figure 2, the CHGNet minimum structure has both high DFT forces and a higher energy than the similar structure suggested by MACE-MP0. Moreover, MACE-MPA undervalues this structure compared to its own suggested GM (purple) on the order 0.2 eV. When relaxed and evaluated at the DFT level, this highly symmetric and well coordinated MACE-MPA-GM structure is slightly less stable than both the DFT-GM and the CHGNet/MACE-MP0-GM structures. On the contrary, when described at the CHGNet and MACE-MP0 level, its stability is strongly underestimated compared to the same GM structures. Furthermore, MACE-MPA evaluates the DFT-GM, CHGNet/MACE-MP0-GM, and red structures as being almost energetically degenerate, a behaviour absent from the other potentials and DFT. The MACE-MPA prediction for the pink and cyan structures is similarly worse than MACE-MP0.

It seems as though the inclusion of the Alexandria dataset codes for the stability of such structures as the MACE-MPA-GM prediction, and highlights the importance of more diverse datasets in capturing the behavior of such phases. However, it is clear that adding more data (even data which one might conclude is more relevant to the system in hand) is not guaranteed to improve or even maintain performance, with MACE-MPA having become more inconsistent on a majority of relative energy predictions than MPtrj only models.

However, the impact of this larger dataset is the contrary when considering $[Ag_2S]_{10}$, for which MACE-MPA alone is able to reproduce the correct DFT-GM structure, cf. Fig. 2. There is demonstrable improvement in mean max forces and the difference in energy between the potential minima and the DFT minima as the dataset size increases, but also evidence that adding more data can change the performance in unexpected ways for systems which were previously consistent. Thus, it is a reasonable assumption to make that when applying such potentials to a problem, particularly one for which the ordering of low energy structures is relevant, one should employ corrective measures.

As a final note, we return to the largest cluster considered in Fig. 2, the $[Ag_2S]_{12}$. For this system the performance of each uMLIP diverges and produce structures which clearly obey some rationale of chemical understanding, but still deviate to the order of at least 1 eV from the DFT-GM, which is a highly symmetric combination of motifs observed in the smaller clusters. In this limit, any of the selected potentials would be unsuitable, and all would require correction.

A. Finding DFT-GMs from active learning of Δ-model corrected uMLIP

In this section, we investigate the ease at which the DFT-GM can be found while performing active learning to correct a uMLIP at the same time as the global optimization is conducted. The five sizes of $[Ag_2S]_X$ clusters present in Fig. 2 are considered, and for each of them, the different search algorithms presented in section II C are employed. For the uMLIP, MACE-MP0 is used throughout, and the discussion of how different uMLIPs perform is deferred to section III B. For the smaller clusters, MACE-MP0 already codes for the correct DFT-GM and requires no Δ-model correction, but we include these cases for completeness.

Figure 4 compares the four different structure search methodologies previously outlined, when applied to the five $[Ag_2S]_X$ cluster sizes. Firstly, $[Ag_2S]_4$ is a very simple problem to solve configurationally, which can be understood intuitively from the relatively small number of atoms present. Furthermore, MACE-MP0 encodes sufficient information to solve the problem correctly (see Fig. 2), so the impact of active learning on the outcome is likely minimal. Figure 4 evidences this, with all but RSS solving the problem almost immediately. Even in this very simple regime, RSS is far from matching the other methods, succeeding in only 80% of cases after a substantial investment of resources.

The struggle of RSS continues for $[Ag_2S]_6$, where once again MACE-MP0 is able to predict the solution correctly itself (cf. Fig. 2). Here, RSS has a ∼15% chance of finding the solution given 48 hours of calculation time (480 CPU hours), compared to the near 100% chance in a fraction of the time for the other methods. For this cluster size, it further becomes apparent that the proposed REX methodology outperforms both GOFEE and Basin Hopping.

$[Ag_2S]_8$ is the point where the configurational complexity and the incomplete map of the PES provided by the uMLIP begin to differentiate the approaches. RSS fails to uncover the true structure even once across all repeats, indicating that attempting relaxation of almost purely random structures as a strategy for identifying the minimum of the PES whilst simultaneously correcting a surrogate potential becomes an unreliable strategy quite rapidly. This highlights the importance of the search strategy, and demonstrates that configurational

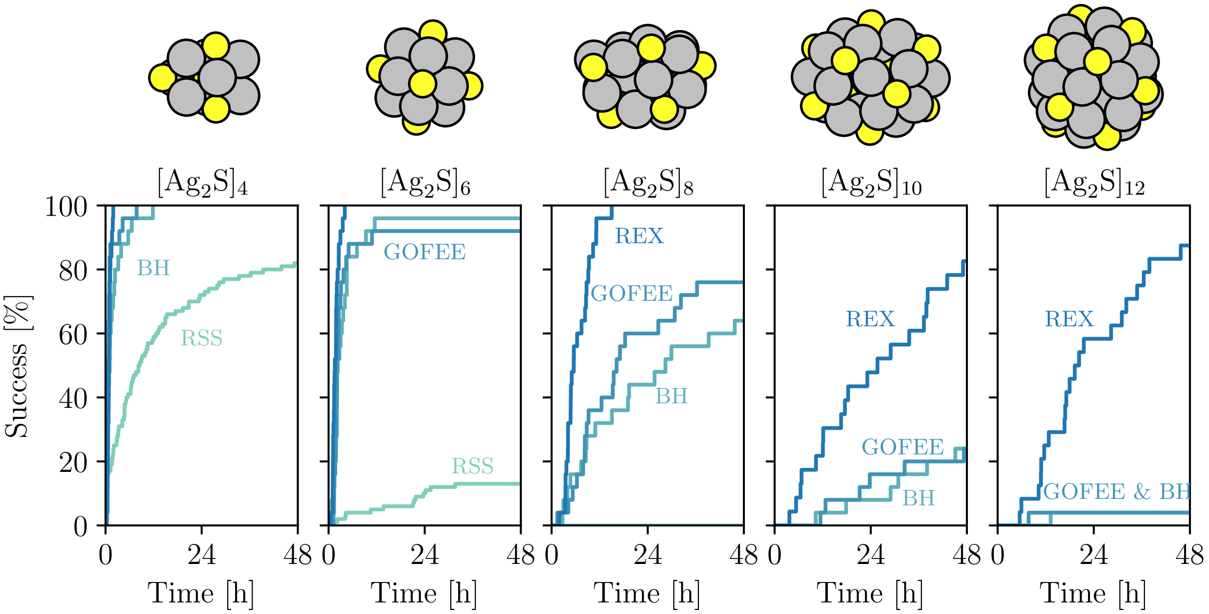

FIG. 4. Global optimization of [Ag$_2$S]$_X$ clusters using active Δ-learning with the MACE-MP0 universal potential and employing the four search methods outlined in Section II C. For each search 25 (100 for RSS) independent repeats were conducted and the success curves report the accumulated share of repeats that have found the GM as a function of elapsed time. The finding of the GM is determined according to a strict spectral graph decomposition.

complexity correlates with the system size, when comparing [Ag$_2$S]$_X$ with $X \in \{4, 6, 8\}$. RSS is omitted from further examples with $X > 8$ on these grounds. REX, however, demonstrates 100% success in about one third of the 48 hours provided, thereby outperforming GOFEE and Basin Hopping, which reach ∼ 75% and ∼ 65% in the full duration, respectively.

Similarly for [Ag$_2$S]$_{10}$, the performance of GOFEE and Basin Hopping drop with respect to REX, which while slower than for the smaller stoichiometry still achieves a success of 80% in the allotted time.

The same trend is echoed once again for [Ag$_2$S]$_{12}$, with diminishing success for the GOFEE and Basin Hopping, and > 80% for REX. Now, only one in 25 Basin Hopping or GOFEE repeats are able to obtain success at this size. The single early success suggests a coincidental occurrence of the solution, indicating that the exploration becomes stunted as time progresses. Meanwhile, REX continues to perform without the same abrupt decrease in success. This demonstrates that REX is a powerful explorative tool, effective at increasing systems sizes and regardless of the understanding of the underlying universal potential. REX effectively couples augmentative active learning to configurational exploration.

Notably, in the regime where *ab initio* evaluation dominates the computational resource budget (many atoms, for periodic systems many **k**-points, expensive XC functionals, and wavefunction methods), REX is at a further advantage when compared to the other methods. Both Basin Hopping and GOFEE devote a larger fraction of their compute budget to performing DFT, and thus perform more DFT calculations per unit time than REX. This can be seen in Supplementary Figure 1, which depicts the same data as Figure 4, repostulated in terms of the number of first-principles calculations performed. This alternative metric more heavily favours REX, as the number of DFT calculations is agnostic to the extent of the repeated replica exchange cycle, and is indicative that REX would also perform excellently in the limit that the computational cost of first principles evaluation is the dominant factor.

B. Comparing the success of different uMLIPs

The previous section demonstrated for MACE-MP0 that the DFT-GMs can reliably be determined in an active learning Δ-model approach. In this section, we widen the investigation to the two other uMLIPs introduced. Since we have already seen the limitations of the optimization algorithms we limit the study to a few combinations of algorithms and cluster sizes.

Figure 5 presents the results. Starting with the RSS algorithm, the smallest cluster, [Ag$_2$S]$_4$, is considered. All three uMLIPs support the finding of the DFT-GM very well with this method. In particular, the CHGNet,

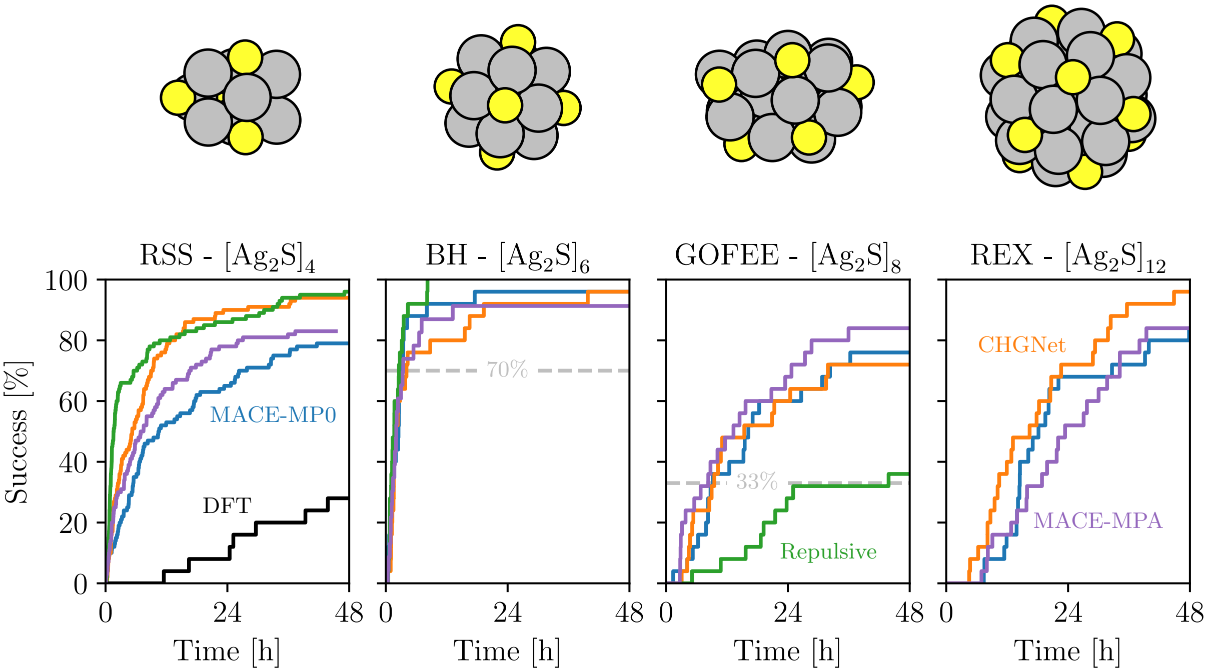

FIG. 5. Success curves for global optimization of $[Ag_2S]_X$ using various combinations of optimization method (left to right) and uMLIP potentials (orange, blue, purple). Included are results using no uMLIP but only a repulsive prior (green) and omitting a surrogate potential altogether (black).

which has the wrong GM encoded, cf. Fig. 2, appears efficient, which we attribute to the fast inference times of that uMLIP.

That speed is more important than precise insight becomes apparent when omitting the uMLIP and instead using a simple repulsive prior [70] that only acts to avoid the collapse of atoms onto the same positions. The success curve of RSS with a repulsive prior is included in green in Fig. 5 and is seen to reach success in e.g. half of the repeats in a matter of 1-2 hours (10-20 CPU hours), for which the uMLIP assisted methods require 5-10 hours (50-100 CPU hours). That a simple potential is useful for solving such problems is consistent with findings of other investigations [14].

Another testimony to the importance of speed of the underlying relaxation calculations comes from the black curve in RSS Fig. 5. The curve shows the success obtained from a purely DFT-based RSS search for the $[Ag_2S]_4$ cluster. This search has no uMLIP and does not construct a surrogate potential to enable cheap structural relaxations, rather every relaxation step is done at the full DFT level. This example represents a historical benchmark moreso than a viable modern strategy.

The Basin Hopping algorithm is applied to the $[Ag_2S]_6$ nanocluster. The rate at which each tested potential reaches 70% success is indistinguishable, with small deviations occurring past this point. Interestingly, the simple repulsive potential is the only one to reach 100% during the allocated time, which as for $[Ag_2S]_4$ highlights the importance of the inference time of the surrogate potential for such problems.

The next optimization algorithm considered in Fig. 5 is GOFEE, which is tested on $[Ag_2S]_8$. The rate at which the DFT-GM is identified starts out very similar for the three uMLIPs. After about 10 hours (100 CPU hours) one in three repeats have indeed found the DFT-GM, irrespective of which uMLIP is used. From there on, the performance of the uMLIPs remains more or less consistent, with MACE-MPA performing the best, reaching over 80% success in the allotted time.

Like for RSS and Basin Hopping, GOFEE can be run without a uMLIP, using instead a repulsive prior in the surrogate potential as originally conceived [25]. This results in the green curve for $[Ag_2S]_8$ in Fig. 5, but unlike for RSS and Basin Hopping, this is now far less efficient than when a uMLIP is available, and requires the full 48 hours (480 CPU hours) to just about reach the 33 % success rate. That the use of the uMLIP is more efficient, we attribute to the increased complexity of the energy landscape for the $[Ag_2S]_8$ compared to those for the smaller clusters solved with RSS and Basin Hopping, and we conjecture that despite its faster inference time, the repulsive prior-based potential suffers from requiring more datapoints for a reliable description of the energy landscape.

Figure 5 finally presents the results of search for the DFT-GM with the REX method for the $[Ag_2S]_{12}$ cluster. As the REX method relies heavily on many rattle-relaxation cycles of walkers, it benefits from the use of the faster CHGNet method when evaluated with a time comparison metric. Notably, even at this size, REX obtains a high success rate of $> 80\%$ regardless of the choice of uMLIP.

We note that it is not possible to use the REX method just based on a repulsive prior, as the extensive searches done by the REX method before any Δ-model can be established would find that atoms should separate as far as possible. This could be circumvented by using a prior with some atomic attraction or by applying pretraining based on precalculated data.

Summarizing this section, we find surprisingly little variance in the efficacy of the three different uMLIPs considered. Despite their differences in initial accuracy, they all support the usage as initial surrogate models and lend themselves to being corrected via our active learning Δ-model protocol.

C. AgS Surface reconstructions

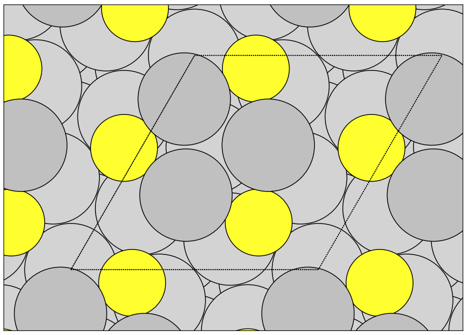

FIG. 6. Depicted is the DFT GM structure for the $(\sqrt{7} \times \sqrt{7})$ surface reconstruction of Ag(111) under sulfurization. Silver atoms which do not reorganize have their color lightened to highlight the reconstruction.

We now move to consider sulfur-induced surface reconstructions. Based on the experimental information available [81, 82], the two systems, Ag(111)$-(\sqrt{7} \times \sqrt{7})$-$Ag_3S_3$ and Ag(100)$-(\sqrt{17} \times \sqrt{17})$-$Ag_{12}S_8$, were chosen. With their different sizes and hence complexity of the involved GMs, these two systems allow for an assessment of the efficiency of the presented methods for active learning and global optimization. The GM structure for the two systems are shown in Figs. 6 and 7, respectively. For the smaller system, all three uMLIPs and the full DFT description agree on a GM in which a triangular Ag_3S_3 motif forms atop the Ag(111). This motif is reminiscent of some of the facets found on the $[Ag_2S]_X$ clusters. The structure shown as the DFT GM in Fig. 6 agrees fully with that proposed in Ref. 81. Rotating the Ag_3S_3 motif by 60 degrees brings the Ag atoms from FCC to HCP sites, and is associated with a small ∼0.01 eV energy penalty. In our statistical analysis below, we therefore do not discriminate between these two solutions.

For the larger system, none of the uMLIPs predict the correct GM as given by DFT. This is detailed in Figure 7 from which it appears that CHGNet and MACE-MP0 energetically overvalue the formation of the Ag_3S_3 motif indicated by the blue triangle (the presence of this motif is unsurprising, given its prevalence in examples thus far). This overvaluing draws the required silver atom away from the slanted Ag_4S_4 motif indicated by the black rhombus. MACE-MPA appears aware that forming the triangles is not as much of an energetic priority, but as seen in the black rhombus, still fails to correctly establish the Ag_4S_4 motif. This failure is reflected by the slight clockwise rotation of the structure with respect to the surface present in the DFT solution i.e. silver atoms in the reconstruction do not fill exact sites of the lattice (hollow), but rather shifted sites. Comparing the DFT-GM of Fig. 7 to that suggested in Ref. 82 leads us to believe that the present search has revealed a new, more stable PBE DFT-GM than hitherto proposed, which further testifies to the efficiency of the REX method in combination with active learning on top of a uMLIP.

Clearly each of the tested uMLIPs has remarkable understanding of the general chemistry involved, and the nature of the required correction, whilst variable between potentials, is small.

Figure 8 presents the results when the GMs for the surface reconstructions are searched with active learning using the different uMLIPs. The first panel of Fig. 8 shows the case of the small, Ag(111)$-(\sqrt{7} \times \sqrt{7})$-$Ag_3S_3$ problem, where the uMLIP all code for the right GM. Owing to the smallness of the problem, we employ the simple RSS optimization method. The success curves show that the GM may be found with 33 % success, i.e. in every 3rd repeat, after about 14 hours (140 CPU hours) in the two MACE potentials, with a considerably slower timeframe for CHGNet. Contrasting to this success, both a set of 75 full DFT random structure searches, and 25 repulsive prior GPR enhanced RSS searches, obtained the solution to the problem only once each in the provided time. We attribute this substantial decrease in performance to the increase in cost of DFT for surface problem when compared to nanoclusters. Clearly the uMLIP Δ-model enhanced RSS enabled this search to succeed where more traditionally it would have been very expensive.

The active learning results for the larger system of Ag(100)$-(\sqrt{17} \times \sqrt{17})$-$Ag_{12}S_8$ are shown in the second panel of Fig. 8. Due to the increased complexity of the problem, the REX optimization method is used. Despite the uMLIPs being confused about the exact GM (cf. Fig. 7), the correct DFT GM is eventually found reliably with all three uMLIPs using the Δ-model approach.

For this system reasonable differences in the efficacy of using the various uMLIPs are found. CHGNet reaches

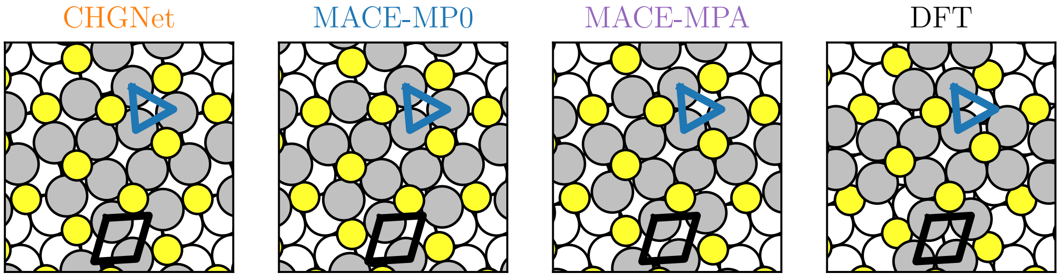

FIG. 7. Structural diagrams of the sulfur-induced surface reconstruction in the Ag(100)-($\sqrt{17} \times \sqrt{17}$) unit cell. From left to right, the minima are obtained through REX searches in CHGNet, MACE-MP0, and MACE-MPA, without active learning. The DFT GM is obtained through REX searches with MACE-MP0 as a prior. The black rhombus present in each panel shows the positions of four particular silver atoms in the DFT solution. The blue triangle highlights three atoms in the MACE-MP0 solution.

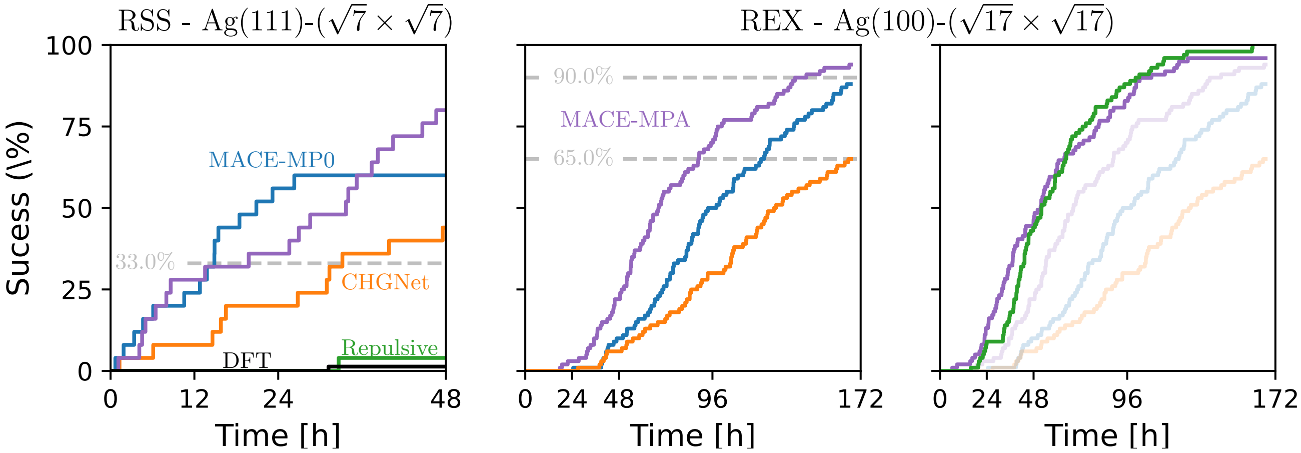

FIG. 8. Success curves for (left) the RSS searches for Ag(111)$-(\sqrt{7} \times \sqrt{7})$-$Ag_3S_3$ and (middle, right) for the REX searches for Ag(100)$-(\sqrt{17} \times \sqrt{17})$-$Ag_{12}S_8$. The curves are colored orange, blue, and purple according to the universal potential used, green for repulsive prior instead of a universal potential, and black for DFT-only omitting the surrogate potential altogether. In the right panel, results are shown for two pretrained potentials: a Δ-model corrected MACE-MPA, which has been provided with initial data (shown in purple), and a GPR model with a simple repulsive prior which has been provided the same data (shown in green). The curves from the previous panel are translucently overlaid for reference. Success is evaluated according to an energy threshold.

65% success in 172 hours (1720 CPU hours), whereas the two MACE models reach around 90% in the same timeframe.

MACE-MPA is found to be considerably faster than the other two uMLIPS at obtaining the solution in general, requiring just over 90 hours to reach the same success as CHGNet did in the full 172. That MACE-MPA is the one to perform the best can be rationalized in terms of its GM deviating less from the DFT-GM than the other two uMLIPs, as discussed in connection with Fig. 7. The Δ-model thus requires less data to be gathered by the active learning in order to correct the uMLIP.

Unlike the nanocluster example, none of the uMLIPs fully encode for the solution. This is not particularly surprising, given that the difference between the DFT solution and the uMLIP solutions boils down to a small rotation of the reconstruction with respect to the underlying surface layer, a highly specific surface phenomenon.

To create such an example where one might be able to compare between a uMLIP which is aware of the solution and one which is not, we construct a MACE-MPA Δ-model from the structures gathered by a REX search for Ag(100)$-(\sqrt{17} \times \sqrt{17})$-$Ag_{12}S_8$. We select the search which contained the lowest energy structure, and train the Δ-model on all 200 structures identified by that search. This model has thus been informed of the relevant surface phenomena, and serves as a point of comparison to the uncorrected uMLIP. Using this Δ-model as the starting point for new REX repeats we obtain the purple curve in the right panel of Fig. 8. This success curve turns out to be a substantial improvement to that of MACE-MPA without pretraining (cf. purple curve in middle panel of Fig. 8), reducing the average time to reach the solution by up to a full day, with the first repeat to find the solution requiring under 6 hours, compared the 18 hours of the base uMLIP. Regardless, the induction time present before the majority the repeats identify the correct solution suggests that the solution is configurationally difficult to come by, requiring a reasonable amount of REX relaxations and swaps to manifest. The most important conclusion to draw is, however, that the magnitude of the improvement achieved with this small amount of extra, highly relevant data, is significant. It then seems likely that the data gathered by the uMLIP REX is particularly descriptive.

To confirm that the data provided to the MACE-MPA Δ-model was indeed efficient at describing the relevant regions of the PES, we repeated the search with a GPR trained on the same dataset, but with the simple repulsive prior instead of the uMLIP. The green curve in the last panel of Fig. 8 shows the success curve of this model, which having seen both the solution and a significant amount of relevant data performs as well as the corrected MACE-MPA from the previous analysis. This is not altogether surprising. It stands to reason that a naive model without extensive extrapolation can produce highly accurate results within the interpolation domain of its training data. This is compelling evidence that the data collected by this method of active learning captures the relevant PES almost completely. With this we underscore that REX results in both robust ab initio quality global optimization and comprehensive sampling of the PES.

IV. CONCLUSION

In this article, we have presented universal potential enhanced structure searching, comparing and contrasting several methods across a variety of silver sulfide nanocluster stoichiometries and surface reconstructions.

Active learning appears a stable method for gathering data for improving the behaviour of uMLIPs. From the four search algorithms considered as vehicles for active learning, REX is the clear standout. We demonstrate significant advantages in efficiency and speed that can be achieved when searching with the REX methodology, combining universal potential Δ-models with replica exchange search. Regardless of the underlying knowledge (or lack thereof) of a particular potential, the REX Δ-model approach is quickly and efficiently able to correct inconsistencies and converge to the DFT solution in comparable timeframes. These results encourage that for any universal potential, such an explorative strategy provides a robust and reliable method to (a) safely apply and (b) improve and discover, with such potentials. Our findings heavily encourage the adoption of such potentials into structure search methodologies, under the provision that they can be corrected incrementally with active learning. Universal potential Δ-models resulting from such searches, conversely, prove to become more effective as search landscapes than the unaugmented potentials, confirming that iterative adaptations effectively alter the form of the PES.

V. DATA AVAILABILITY

The datasets supporting the findings in this paper are available on Zenodo at `https://doi.org/10.5281/zenodo.16368472` along with Python scripts to reproduce our findings. The optimization methods employed are available in version 3.10.2 of the AGOX package available to install from Pypi and source code is openly available on Gitlab `https://gitlab.com/agox/agox`.

VI. ACKNOWLEDGEMENTS

We acknowledge support from VILLUM FONDEN through Investigator grant, project no. 16562, and by the Danish National Research Foundation through the Center of Excellence "InterCat" (Grant agreement no: DNRF150).

[1] K. Butler, D. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature **559**, 547 (2018).

[2] O. von Lilienfeld and K. Burke, Nature Communications **11**, 4895 (2020).

[3] N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. S. Smith, B. Nebgen, R. A. Messerly, Y. W. Li, A. Boldyrev, K. Barros, O. Isayev, and S. Tretiak, Nature Reviews Chemistry **6**, 653 (2022).

[4] J. Xia, Y. Zhang, and B. Jiang, Chem. Soc. Rev. **54**, 4790 (2025).

[5] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[6] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[7] L. Hörmann, W. G. Stark, and R. J. Maurer, npj Computational Materials **11**, 196 (2025).

[8] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).

[9] Y. Mishin, Acta Materialia **214**, 116980 (2021).

[10] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X **8**, 041048 (2018).

[11] V. Kapil, C. Schran, A. Zen, J. Chen, C. J. Pickard, and A. Michaelides, Nature **609**, 512 (2022).

[12] T. Maxson, A. Soyemi, B. W. J. Chen, and T. Szilvási, The Journal of Physical Chemistry C **128**, 6524 – 6537

(2024).

[13] V. L. Deringer, C. J. Pickard, and G. Csányi, Phys. Rev. Lett. **120**, 156001 (2018).

[14] C. J. Pickard, Phys. Rev. B **106**, 014102 (2022).

[15] N. Rønne, M.-P. V. Christiansen, A. M. Slavensky, Z. Tang, F. Brix, M. E. Pedersen, M. K. Bisbo, and B. Hammer, J. Chem. Phys. **157**, 174115 (2022).

[16] C. Larsen, S. Kaappa, A. Vishart, T. Bligaard, and K. W. Jacobsen, npj Computational Materials **11**, 222 (2025).

[17] Z. Tang, S. T. Bromley, and B. Hammer, J. Chem. Phys. **158**, 224108 (2023).

[18] A. Khorshidi and A. A. Peterson, Computer Physics Communications **207**, 310 (2016).

[19] J. Timmermann, F. Kraushofer, N. Resch, P. Li, Y. Wang, Z. Mao, M. Riva, Y. Lee, C. Staacke, M. Schmid, C. Scheurer, G. S. Parkinson, U. Diebold, and K. Reuter, Phys. Rev. Lett. **125**, 206101 (2020).

[20] M. J. Waters and J. M. Rondinelli, Journal of Physics: Condensed Matter **34**, 385901 (2022).

[21] S. Ma, C. Shang, and Z.-P. Liu, The Journal of Chemical Physics **151**, 050901 (2019).

[22] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, Phys. Rev. Mater. **3**, 023804 (2019).

[23] N. Bernstein, G. Csányi, and V. Deringer, npj Computational Materials **5**, 99 (2019).

[24] R. Jinnouchi, F. Karsai, and G. Kresse, Phys. Rev. B **100**, 014105 (2019).

[25] M. K. Bisbo and B. Hammer, Phys. Rev. Lett. **124**, 086102 (2020).

[26] M. L. Paleico and J. Behler, J. Chem. Phys. **153**, 054704 (2020).

[27] J. Wang, H. Gao, Y. Han, C. Ding, S. Pan, Y. Wang, Q. Jia, H.-T. Wang, D. Xing, and J. Sun, Nat. Sci. Rev. **10**, nwad128 (2023).

[28] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, and S. Tretiak, Chemical Reviews **124**, 13681 (2024).

[29] R. Wanzenböck, E. Heid, M. Riva, G. Franceschi, A. M. Imre, J. Carrete, U. Diebold, and G. K. H. Madsen, Digital Discovery **3**, 2137 (2024).

[30] Y. Lee, X. Chen, S. M. Gericke, M. Li, D. N. Zakharov, A. R. Head, J. C. Yang, and A. N. Alexandrova, Angewandte Chemie International Edition **64**, e202501017 (2025).

[31] F. Grasselli, S. Chong, V. Kapil, S. Bonfanti, and K. Rossi, Digital Discovery , Accepted Manuscript (2025).

[32] V. G. Satorras, E. Hoogeboom, and M. Welling, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021) pp. 9323–9332.

[33] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nat Commun **13**, 2453 (2022).

[34] N. Leimeroth, L. C. Erhard, K. Albe, and J. Rohrer, “Machine-learning interatomic potentials from a users perspective: A comparison of accuracy, speed and data efficiency,” (2025), arXiv:2505.02503 [cond-mat.mtrl-sci].

[35] C. Chen and S. P. Ong, Nature Computational Science **2**, 718 (2022).

[36] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Nat. Mach. Intell. **5**, 1031 (2023).

[37] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O’Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, “A foundation model for atomistic materials chemistry,” (2024), arXiv:2401.00096 [physics.chem-ph].

[38] N. T. Taylor, J. Pitfield, F. H. Davies, and S. P. Hepplestone, arXiv preprint arXiv:2504.02528 (2025).

[39] K. S. Jakob, K. Reuter, and J. T. Margraf, Advanced Intelligent Discovery , 202500031 (2025).

[40] B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson, and G. Ceder, npj Computational Materials **11**, 9 (2025).

[41] M. M. Ghahremanpour, P. J. Van Maaren, and D. Van Der Spoel, Scientific Data **5**, 180062 (2018).

[42] J. Riebesell, R. E. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain, and K. A. Persson, Nature Machine Intelligence **7**, 836 (2025).

[43] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, arXiv preprint arXiv:2502.12147 (2025).

[44] B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. R. Kitchin, D. S. Levine, *et al.*, arXiv preprint arXiv:2506.23971 (2025).

[45] B. Focassio, L. P. M. Freitas, and G. R. Schleder, ACS Applied Materials & Interfaces **17**, 13111 (2024).

[46] D. Marchand, MRS Bulletin **50**, 805 (2025).

[47] B. Póta, P. Ahlawat, G. Csányi, and M. Simoncelli, arXiv preprint arXiv:2408.00755 (2024).

[48] H. Lee, V. I. Hegde, C. Wolverton, and Y. Xia, Materials Today Physics **53**, 101688 (2025).

[49] H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides, and V. Kapil, Faraday Discussions **256**, 120 (2025).

[50] J. Pitfield, F. Brix, Z. Tang, A. M. Slavensky, N. Rønne, M.-P. V. Christiansen, and B. Hammer, Phys. Rev. Lett. **134**, 056201 (2025).

[51] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).

[52] A. V. Terentjev, L. A. Constantin, and J. M. Pitarke, Physical Review B **98**, 214108 (2018).

[53] J. Sun, A. Ruzsinszky, and J. P. Perdew, Physical Review Letters **115**, 036402 (2015).

[54] A. Nandi, C. Qu, P. L. Houston, R. Conte, and J. M. Bowman, J. Chem. Phys **154**, 051102 (2021).

[55] W. Yang and P. W. Ayers, arXiv preprint arXiv:2403.04604 (2024).

[56] P. Bandyopadhyay, B. K. Isamura, and P. L. Popelier, The Journal of Chemical Physics **162**, 074102 (2025).

[57] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).

[58] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[59] M.-P. V. Christiansen, N. Rønne, and B. Hammer, Machine Learning: Science and Technology **5**, 045029 (2024).

[60] P. Lyngby, C. Larsen, and K. W. Jacobsen, Physical Review Materials **8**, 123802 (2024).

[61] M.-P. V. Christiansen and B. Hammer, The Journal of Chemical Physics **162**, 184701 (2025).

[62] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Computer Physics Communications **247**, 106949 (2020).

[63] J. Laakso, L. Himanen, H. Homm, E. V. Morooka, M. O. Jäger, M. Todorović, and P. Rinke, The Journal of Chemical Physics **158**, 234802 (2023).

[64] M.-P. V. Christiansen, N. Rønne, and B. Hammer, The Journal of Chemical Physics **157**, 054701 (2022).

[65] R. Fletcher, *Practical methods of optimization* (John Wiley & Sons, 2000).

[66] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, Journal of Physics: Condensed Matter **29**, 273002 (2017).

[67] J. J. Mortensen, A. H. Larsen, M. Kuisma, A. V. Ivanov, A. Taghizadeh, A. Peterson, A. Haldar, A. O. Dohn, C. Schäfer, E. Ö. Jónsson, *et al.*, The Journal of Chemical Physics **160**, 092503 (2024).

[68] C. J. Pickard and R. Needs, Journal of Physics: Condensed Matter **23**, 053201 (2011).

[69] N. Metropolis and S. Ulam, Journal of the American Statistical Association **44**, 335 (1949).

[70] M. K. Bisbo and B. Hammer, Physical Review B **105**, 245404 (2022).

[71] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).

[72] R. Zhou and B. J. Berne, Proceedings of the National Academy of Sciences **99**, 12777 (2002).

[73] A. E. García and J. N. Onuchic, Proceedings of the National Academy of Sciences **100**, 13898 (2003).

[74] R. Yamamoto and W. Kob, Physical Review E **61**, 5473 (2000).

[75] R. H. Swendsen and J.-S. Wang, Physical Review Letters **57**, 2607 (1986).

[76] N. Unglert, L. B. Pártay, and G. K. Madsen, arXiv preprint arXiv:2505.04390 (2025).

[77] U. H. Hansmann, Chemical Physics Letters **281**, 140 (1997).

[78] P. A. Frantsuzov and V. A. Mandelshtam, Physical Review E **72**, 037102 (2005).

[79] C. Song and Z. Tian, Journal of Molecular Modeling **25**, 310 (2019).

[80] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, The Journal of Chemical Physics **132**, 154104 (2010).

[81] M. Shen, D.-J. Liu, C. J. Jenks, and P. A. Thiel, The Journal of Physical Chemistry C **112**, 4281 (2008).

[82] S. M. Russell, M. Shen, D.-J. Liu, and P. A. Thiel, Surface Science **605**, 520 (2011).