

Hyperactive Learning (HAL) for Data-Driven Interatomic Potentials

Oord, Cas van der; Sachs, Matthias; Kovács, Dávid Péter; Ortner, Christoph; Csányi, Gábor

DOI:

[10.48550/arXiv.2210.04225](https://doi.org/10.48550/arXiv.2210.04225)

License:

None: All rights reserved

Document Version

Publisher's PDF, also known as Version of record

Citation for published version (Harvard):

Oord, CVD, Sachs, M, Kovács, DP, Ortner, C & Csányi, G 2022 'Hyperactive Learning (HAL) for Data-Driven Interatomic Potentials'. <https://doi.org/10.48550/arXiv.2210.04225>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Hyperactive Learning (HAL) for Data-Driven Interatomic Potentials

Cas van der Oord,* Dávid Péter Kovács, and Gábor Csányi
University of Cambridge, Cambridge, CB2 1PZ, U.K.

Matthias Sachs
University of Birmingham, Birmingham, B15 2TT, U.K.

Christoph Ortner
University of British Columbia, Vancouver, BC, V6T 1Z2, Canada
(Dated: October 11, 2022)

Data-driven interatomic potentials have emerged as a powerful class of surrogate models for *ab initio* potential energy surfaces that are able to reliably predict macroscopic properties with experimental accuracy. In generating accurate and transferable potentials the most time-consuming and arguably most important task is generating the training set, which still requires significant expert user input. To accelerate this process, this work presents *hyperactive learning* (HAL), a framework for formulating an accelerated sampling algorithm specifically for the task of training database generation. The overarching idea is to start from a physically motivated sampler (e.g., molecular dynamics) and a biasing term that drives the system towards high uncertainty and thus to unseen training configurations. Building on this framework, general protocols for building training databases for alloys and polymers leveraging the HAL framework will be presented. For alloys, fast ($<100 \mu\text{s}/\text{atom}/\text{cpu-core}$) ACE potentials for AlSi10 are created that able to predict the melting temperature with good accuracy by fitting to a minimal HAL-generated database containing 88 configurations (32 atoms each) in 17 seconds using 8 cpu threads. For polymers, a HAL database is built using ACE able to determine the density of a long polyethylene glycol (PEG) polymer formed of 200 monomer units with experimental accuracy by only fitting to small isolated PEG polymers with sizes ranging from 2 to 32.

I. INTRODUCTION

Over the last decade there has been rapid progress in the development of data-driven interatomic potentials, see the review papers [1–6]. Many systems are often too complex to be modelled by an empirical description yet inaccessible to electronic structure methods due to prohibitive computational cost. Richly parametrised data-driven interatomic potentials bridge this gap and are able to successfully describe the underlying chemistry and physics by approximating the potential energy surface (PES) with quantum mechanical accuracy [7–9]. This approximation is done by regressing a *high-dimensional* model to training data collected from electronic structure calculations.

Over the years many approaches have been explored using a range of different model architectures. These include Artificial Neural Networks (ANN) based on atom centered symmetry functions[10] and have been used in models such as ANI[11, 12] and DeepMD [13]. Another widely used approach is Gaussian Process Regression (GPR) implemented in models such as SOAP/GAP [14, 15], FCHL [16] and sGDML [17]. Linear approximations of the PES have also been introduced initially by using permutation invariant polynomials (PIPs) [18] and the more recent atomic PIPs variant[19, 20]. Other linear models include spectral neighbour analysis potentials [21] based on the bispectrum [22], moment tensor potentials [23] and the atomic cluster expansion (ACE) [24–26]. More recently, message passing neural network (MPNN) architectures have been introduced [27–34] the most recent of which have been able

* casv2@cam.ac.uk

to outperform any of the previously mentioned models regarding accuracy on benchmarks such as MD17 [35] and ISO17 [36]. Central to all of these models is that they are fitted to a training database comprised of configurations R labelled with observations comprising total energy \mathcal{E}_R , forces \mathcal{F}_R and perhaps virial stresses \mathcal{V}_R , obtained from electronic structure simulations. By performing a regression on the training data model predictions E of the total energy, and estimates of the respective forces $F_i = -\nabla_i E$ can be determined. Here, the ∇_i operator denotes the gradient with respect to the position of atom i .

Building suitable training databases remains a challenge and the most time consuming task in developing general data-driven interatomic potentials [37–39]. Databases such as MD17 and ISO17 are typically created by performing Molecular Dynamics (MD) simulations on the structures of interest and selecting decorrelated configurations along the trajectory. This approach samples the potential energy surface according to its Boltzmann distribution. Once the training database contains sufficient number of configurations, a high dimensional model may be regressed in order to accurately interpolate its potential energy surface. The interpolation accuracy can be improved by further sampling, albeit with diminishing returns. However, it is by no means clear that the Boltzmann distribution is the optimal measure, or even a “good” measure, from which to draw samples for an ML training database. Indeed, it likely results in severe undersampling of configurations corresponding to defects and transition states, particularly for material systems with high barriers, which nevertheless have a profound effect on material properties and are often the subject of intense study.

A lack of training data in a sub-region can lead to deep unphysical energy minima in trained models, sometimes called “holes”, which are well known to cause catastrophic problems for MD simulations: the trajectory can get trapped in these unphysical minima or even become unstable numerically for normal step sizes. A natural strategy to prevent such problems

is active learning (AL): the simulation is augmented with a stopping criterion aimed at detecting when the model encounters a configuration for which the prediction is unreliable. Intuitively, one can think of such configurations as being “far” from the training set. When this situation occurs, a ground-truth evaluation is triggered, the training database extended, and the model refitted to the enlarged database. In the context of data-driven interatomic potentials, this approach was successfully employed by the linear moment tensor potentials [40, 41] and the Gaussian process (GP) based methods FLARE [42, 43] and GAP [44] which both use site energy uncertainty arising from the GP to formulate a stopping criterion in order to detect unreliable predictions during simulations.

The key contribution of this work is the introduction of the *hyperactive learning* framework. Rather than relying on normal MD to sample the potential energy and wait until an unreliable prediction appears (which may take a very long time once the model is decent), we continually bias the MD simulation towards regions of high uncertainty. By balancing the physical MD driving force with such a bias we accelerate the discovery of unreliably predicted configurations but retain the overall focus on low energy regions that are important for modelling.

Concretely, given some measure of model uncertainty σ , the HAL framework replaces the energy landscape E in an MD simulation with the *HAL potential energy* surface E_{HAL}

$$E_{\text{HAL}} = E - \tau\sigma. \quad (1)$$

This biases the MD towards configurations with high associated uncertainty. The *biasing strength* τ balances the strength of the biasing potential with the physical potential energy E . In order to ensure that configurations along the HAL-MD trajectory remain energetically sensible, some judicious setting of τ is required; on-the-fly auto-tuning of τ is presented in the Methods section. Conceptually, the introduction of a biasing potential that accelerates the exploration of relevant configurations is reminiscent of adaptive biasing strategies such as metadynamics [45, 46], umbrella sampling [47, 48],

and similar methods (e.g., [49, 50]), which were developed to accelerate free-energy computation. While the biasing force in these methods is implicitly specified by the choice of a collective variable, the direction of the biasing force in HAL is the result of the choice of the uncertainty measure σ .

It should also be noted that in balancing low energy against uncertainty, E_{HAL} takes a similar role in our framework as the acquisition function in Bayesian Optimisation (BO) [51]. BO is designed to optimise an expensive black box function by balancing exploration and exploitation [52]. It has been shown to yield state-of-the-art results for optimisation problems while simultaneously minimising incurred computational costs by requiring few evaluations of the black box function [53]. In HAL, the equivalent to the black box function is an ab initio calculation, yet we do not seek to minimise but rather to explore the black box, guided or accelerated by the biasing term.

We make the general HAL concept concrete in the context of the ACE framework [24, 25], however, the methods we propose are immediately applicable to linear models and to Gaussian process type models, and are in principle also extendable to any other ML potential, including deep neural network models. In the context of linear ACE models, described in detail in the Methods section, the site energy is defined as a linear combination of basis functions,

$$E_i = \mathbf{c} \cdot \mathbf{B}_i. \quad (2)$$

and total energy, $E = \sum_i E_i = \mathbf{c} \cdot \mathbf{B}$ where $\mathbf{B} = \sum_i \mathbf{B}_i$.

The prediction of the uncertainty σ can, for example, be obtained through the use of an ensemble. Different methods of setting up such ensembles for linear, GP or NN frameworks can be used, such as dropout [54], or bootstrapping [55]. In this work, we leverage the linearity of the ACE model and adopt a Bayesian view of the regression problem to obtain a rigorous committee: Assuming an isotropic Gaussian prior on the model parameters and Gaussian independent and identically distributed (i.i.d) noise on observations, yields an explicit posterior distribution

$\pi(\mathbf{c})$ of the parameters from which one can deduce the variance σ_E^2 of the posterior-predictive distribution of total energies,

$$\sigma_E^2 = \frac{1}{\lambda} + \mathbf{B}^T \boldsymbol{\Sigma} \mathbf{B}, \quad (3)$$

where the covariance matrix $\boldsymbol{\Sigma}$ is defined as

$$\boldsymbol{\Sigma}^{-1} = \alpha \mathbf{I} + \lambda \boldsymbol{\Psi}^T \boldsymbol{\Psi}. \quad (4)$$

Here, α, λ are hyperparameters whose treatment is detailed in the Methods section, and $\boldsymbol{\Psi}$ is the corresponding design matrix of the linear regression problem and depends on the observations to which the ACE model is fitted.

The evaluation of the uncertainty or variance σ_E^2 in equation (3) is computationally expensive for a large basis \mathbf{B} ; scaling as $O(N_{\text{basis}}^2)$. To improve computational efficiency, σ_E^2 can be approximated by using an ensemble $\{\mathbf{c}^k\}_{k=1}^K$ obtained by sampling from the posterior $\pi(\mathbf{c})$ (see Methods for further details), resulting in

$$\tilde{\sigma}_E^2 = \frac{1}{\lambda} + \frac{1}{K} \sum_{k=1}^K (E^k - \bar{E})^2, \quad (5)$$

where $\bar{E} = \bar{\mathbf{c}} \cdot \mathbf{B}$ with $\bar{\mathbf{c}}$ being the posterior mean of the posterior distribution whose closed form is provided in (22) of the methods section. This is computationally efficient to evaluate, requiring a single basis evaluation \mathbf{B} followed by K dot-products with the ensemble parameters.

Throughout the remainder of this article we will fix the choice of uncertainty measure and its' approximation to $\sigma = \sigma_E$ and $\tilde{\sigma} = \tilde{\sigma}_E$, respectively, in the definition of the HAL energy to be the standard deviation of the posterior-predictive distribution of energy as outline above. From both a theoretical and modelling perspective, it would be of interest to consider other measures of uncertainty as biasing terms. Further discussion of this aspect is provided in the methods section.

Having introduced HAL-MD it remains to specify a stopping criterion that can be used to terminate the dynamics and extract new training configurations. To that end we introduce a *relative force uncertainty*, f_i , which is attractive

from a modelling perspective, as for instance liquid and phonon properties require vastly different absolute force accuracy but similar relative force accuracy, typically on the order of 3-10%. Given the model committee we introduced to define $\tilde{\sigma}$ we define

$$f_i = \frac{\frac{1}{K} \sum_{k=1}^K \|F_i^k - \bar{F}_i\|}{\|\bar{F}_i\| + \varepsilon}, \quad (6)$$

where \bar{F}_i is the mean force prediction. Further, ε is a regularising constant to prevent divergence of the fraction, and to be specified by the user, often set to the typical force magnitude as predicted by $\|\bar{F}\|$. During HAL simulations, f_i provides a computationally efficient means to detect emerging local (force) uncertainties and trigger new ab initio calculations once it exceeds a predefined tolerance,

$$\max_i f_i > f^{\text{tol}}. \quad (7)$$

The specification of f^{tol} is both training data and model specific, and often requires careful tuning to achieve good performance. Too low f^{tol} keeps triggering unnecessary ab initio calculations, whereas too high leads to generation of unphysical high energy configurations. To avoid manual tuning and aid generality, we normalise f_i onto $[0, 1]$ through the application of the softmax function $s(f_i)$, resulting in the new stopping criterion

$$\max_i \frac{\exp f_i}{\sum_i \exp f_i} > s^{\text{tol}}, \quad (8)$$

where we use the default tolerance $s^{\text{tol}} = 0.5$.

The paper is structured as follows. Following an initial discussion of the performance of the relative force error measure f_i , its ability to predict true error is investigated and its performance benchmarked by assembling a reduced diamond structure silicon database. Next, the HAL framework is used to build training databases for an alloy (AlSi10) and polymer (polyethylene glycol or PEG) from scratch and the ability of the resulting ACE models are able to accurately predict the AlSi10 melting temperature and PEG density as shown.

Note that just a few days before our submission to the arXiv preprint server, the preprint [56] appeared and follows a very similar strategy to ours.

II. RESULTS AND DISCUSSION

A. Filtering an existing training set

Before illustrating the HAL algorithm itself, we first demonstrate the ability of the relative force error estimate f_i in Eq. (6) to detect true relative force errors. To that end, we will use the estimator to significantly reduce a large training set while maintaining accurate model properties relative to the DFT reference. The database we use for this demonstration was originally developed for a Si GAP model [38] covering a wide range of structures ranging from bulk crystals in various phases, amorphous, liquid and vacancy configurations. The filtering process builds a reduced database by starting from a single configuration and selecting configurations containing the maximum f_i from the remaining test configurations. Iterating this process accelerates the learning rate and rapidly converges model properties with respect to the DFT reference. The models we train in this was are linear ACE models containing basis functions up to correlation order $\nu=3$, polynomial degree 20, outer cutoff set to 5.5 Å and inner cutoff set to the closest interatomic distance in the training database. An auxiliary pair potential basis was used using polynomial degree 3 and outer cutoff 7.0 Å and no inner cutoff. The weights for the energy w_E , forces w_F and virials w_V , which are described in detail in the Methods section, were set to 5.0/1.0/1.0. The size of the committees used to determine f_i was $K = 32$.

1. Si diamond: error correlation and convergence

Prior to training database reduction the ability of the relative force error estimate f_i to predict relative force error is investigated. Fig. 1a compares the maximum relative force error in

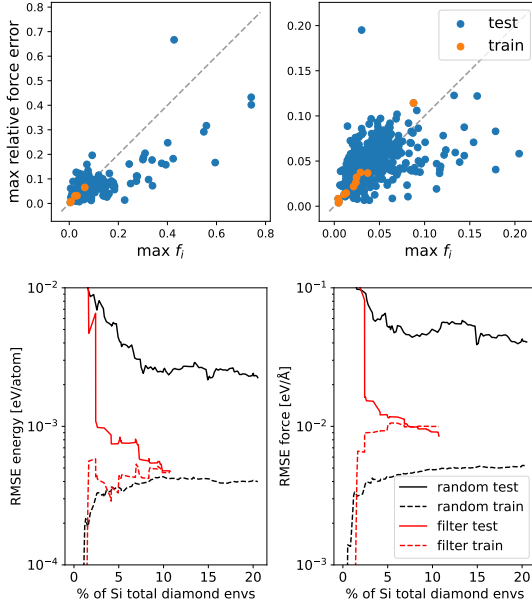


FIG. 1: a) Maximum relative force error estimate $\max f_i$ versus error correlation plots for silicon diamond containing 4 and 10 training configurations. b) Learning rate comparison between filtering and random selection for silicon diamond.

a configuration against the maximum of f_i for two different training databases, containing 4 and 10 silicon diamond configurations respectively. The test configurations are the remaining configurations contained in the 489 silicon diamond configurations as part of the entire silicon database (totalling 16708 local environments). The regularising constant ε was set to the mean force magnitude as predicted by the mean parameterisation. Both figures show good correlation between maximum relative force error and $\max f_i$, therefore making it a suitable criterion to be monitored during (H)AL strategies.

By leveraging the correlation of f_i with true relative force error the existing silicon diamond database can be reduced by iteratively selecting configurations containing the largest relative force uncertainty as part of a greedy al-

gorithms strategy. To demonstrate this a single configuration from the 489 silicon diamond configurations contained in the silicon database was chosen first and fitted. Next, f_i was determined over the remaining configurations and the configuration containing the largest $\max f_i$ added to the training database. This process was repeated and the train and test error of this filtering procedure for silicon diamond is shown in Fig. 1b. It is benchmarked against performing random selection whereby, starting from the same initial configuration, test configurations were chosen at random from the pool of remaining test configurations. The result indicates that f_i accurately detects configurations with large errors and manages to accelerate the learning rate significantly relative to random selection. Good generalisation between training and test errors is achieved by using around 5% of the total environment contained in the original silicon diamond database.

2. Si diamond: property convergence

The significant acceleration of the learning rate shown in Fig. 1b shows that generalisation between train and test error is rapidly achieved, in turn suggesting that property convergence is accelerated too. This is investigated by using the fitted linear ACE models during the filtering process and examining their predicted properties and comparing to the DFT reference. These investigated properties included: elastic constants, energy volume curves, phonon spectrum and thermal properties for bulk silicon diamond. Three linear ACE models were chosen containing 3%, 4% and 5% percent of the original total of silicon diamond environments, effectively slicing Fig. 1b along the x-axis. These models were fitted to 9 configurations (424 environments), 13 configurations (460 environments) and 17 configurations (608 environments) of silicon diamond respectively.

Fig. 2 demonstrates that indeed property convergence for the energy volume curves, phonon spectrum and thermal properties are rapidly achieved by fitting to a fraction of the orig-

	B [GPa]	c_{11} [GPa]	c_{12} [GPa]	c_{44} [GPa]
ACE 3% envs	98.2 [+19%]	188.1 [+28%]	53.3 [+6%]	79.7 [+8%]
ACE 4% envs	84.2 [+2%]	159.8 [+9%]	46.4 [-8%]	75.7 [+3%]
ACE 5% envs	82.5 [0%]	148.7 [+1%]	49.3 [-2%]	73.7 [+1%]
DFT	82.6	147.2	50.3	73.1

TABLE I: Convergence of the elastic moduli (GPa) of the filtered ACE models relative to the CASTEP DFT reference.

inal database. Fitting to 5% of the original database reaches sufficient accuracy to describe all properties with good accuracy with respect to the DFT reference. This is again confirmed by elastic constants as predicted by the respective models as shown in Table. I. The convergence of the phonon spectrum in Fig. 2 is particularly noteworthy as relative errors on the order of a few percent on small forces ~ 0.01 eV/Å are required to be described accurately in order to describe the phonon spectrum well. This is achieved by designing f_i to consider relative force errors, which the greedy algorithm manages to detect and suppresses during the filtering process.

B. AlSi10

This section outlines the general HAL protocol for building training databases for alloys and demonstrates how an AlSi10 linear ACE model is built from scratch in an automated fashion. By using the relative force error estimate f_i previously discussed as a stopping criterion to trigger ab initio evaluations it will be shown how an ACE model is created for AlSi10 using HAL. The HAL generated ACE model will be able to accurately model the liquid-solid phase transition and predict its melting temperature with excellent accuracy. The ACE models used in this section contained basis functions upto $\nu = 2$ and polynomial degree 13 as well as an outer cutoff 5.5 Å. The ACE inner cutoff was set to 1.5 Å during the HAL stage of collecting data and moved towards the closest interatomic distance once all training data had been generated. An auxiliary pair potential V_2 added to

aid stability also added to the basis including functions up to polynomial degree 13 and an outer cutoff of 6.0 Å. The weights for the energy w_E , forces w_F and virials w_V were set to 15.0/1.0/1.0.

The HAL procedure of building ACE models for alloys first creates a random alloy database, before melting the alloy configurations and generating a liquid alloy database. The solid and liquid databases are afterwards combined in order to create a model accurately described both phases. Building the random alloy training database starts off by first creating a small initial random alloy database from which HAL dynamics is started. This initial random alloy database was formed of 32-atom FCC lattice configurations populated with 29 Al and 3 Si atoms, equivalent to 9.7 weight percent Si. The initial random alloy starting database contained 10 configurations with lattice constants ranging from 3.80 Å to 4.04 Å evaluated using DFT. The main parameters chosen for the DFT calculations in this section for CASTEP [58] are as follows: plane-wave cutoff 300 eV, kpoint spacing 0.04 Å⁻¹, 0.1 eV smearing, Pulay density mixing scheme and finite basis correction.

An adaptive biasing parameter $\tau_r=0.05$ was chosen (for explicit definition see Methods section) and the temperature set to $T_{\text{solid}}=800\text{K}$ in order to build the random solid alloy database starting from the 10 initial structures previously described. Besides running biased dynamics, HAL performed cell volume adjusting (adding Gaussian noise to cell vectors) and element swapping Monte Carlo (MC) steps during the simulation on the HAL potential energy surface E_{HAL} . These steps were accepted or rejected according to the Metropolis-Hastings al-

α'	N_{basis}	Performance ($\mu\text{s}/\text{atom}$ /core)	Fit (s)	Training Error		Test Error	
				E (meV/at)	F (eV/Å)	E (meV/at)	F (eV/Å)
1k	38	62	2	7.693	0.135	8.006	0.147
10k	116	83	3	4.199	0.095	6.229	0.104
80k	295	85	17	2.401	0.080	5.131	0.089
300k	621	99	63	1.869	0.074	5.188	0.095

TABLE II: Train/test error splits for HAL generated AlSi10 database for varying ARD tolerance α' . Larger ARD tolerance α' includes more basis functions but increases performance and fitting time.

gorithm [59].

During HAL dynamics the (softmax normalised) relative force estimate s_i is evaluated and a ground-truth evaluation triggered once a predefined tolerance of $s^{\text{tol}}=0.5$ is met. A total of 42 HAL configurations were sampled as the HAL dynamics at this stage was stable for 5000 steps reliably. The pressure P , temperature T and f_i are shown in Fig. 3 for four iterations with the first three being included in the training database, e.g. below or equal to iteration 42. The strong oscillations in the pressure P are due to the volume and element swapping MC steps being accepted. A fourth iteration, referred to as iteration 43, is shown on the far right demonstrating that increasing the biasing strength to $\tau_r=0.10$ accelerates the HAL dynamics to failure more rapidly accelerating the discovery of uncertain configurations exhibiting large relative force errors and resulting model failure.

Increasing the temperature to $T_{\text{liquid}}=3000\text{K}$ a liquid random alloy training database was assembled using the HAL generated random alloy training database as initial starting configurations. No volume or swap MC steps were performed during this simulation, but a barostat was added and set to 0.1 GPa in order to control the pressure during the simulation. After generating generating 46 liquid alloy configurations using HAL the dynamics was stable for 5000 steps reliably and HAL dynamics terminated.

Combining the 42 HAL generated random alloy configurations and 46 HAL generated liquid configurations formed the complete training

database used to create AlSi10 linear ACE models. Using the previously described ACE parameters the design matrix Ψ assembled shaped $N_{\text{obs}} \times N_{\text{basis}}$ was 9064×723 . Fitting was performed using ARD regression with various different thresholds α' and details are shown in Table II. Both the contents and shape of the design matrix Ψ as well as ARD are discussed in the Methods section. Increasing α' lowers the relevance criterion for the linear ACE basis functions in turn decreasing sparsity. A clear trade-off between sparsity and training error can be seen in Table II which also includes model evaluation performance and fitting times. The test set was assembled by continuing HAL iterations for both the random alloy and liquid and contained 14 HAL solid and 14 HAL liquid configurations. Increasing α' not only decreases training error but also test set error up to $\alpha' = 300k$ for which the test set error increases, a sign of overfitting. Due to the relatively small training database size the fitting time remains low, around a minute or less using 8 threads on Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz. Performance testing was done using LAMMPS and the PACE evaluator [60] on the exact same processor and show that all models are within 100 $\mu\text{s}/\text{atom}/\text{core}$ per MD step.

Further analysis of the ARD fitted models was done by examining the absolute value of the coefficients $|c_i|$. ARD performs feature selection by determining the relevance of the basis functions and weighting them accordingly in the regression. Basis functions with a relevance below the predefined threshold are pruned away

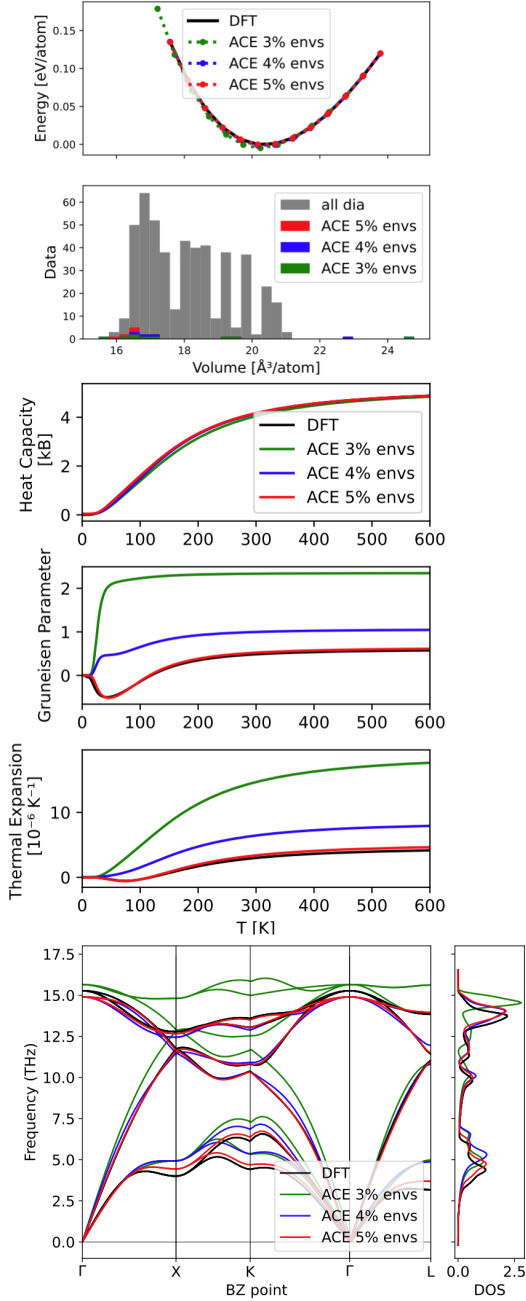


FIG. 2: Property convergence for the energy volume (top), thermal properties (middle) and phonon spectrum (bottom) for filtering silicon diamond ACE models.

as can be seen in Fig. 4. Large coefficients are given to the pair interactions described by the auxiliary basis V_2 and two-body components of the ACE basis for all models, which is intuitive as most binding energy is stored in these pair interactions. Increasing α' results in more (less relevant) basis functions being included with relatively smaller coefficients. For $\alpha' = 300k$ many of these low relevance coefficients of around 10^{-4} are included in the fit indicating a degree of overfitting - as confirmed by the test set error increase in Table II.

Next the melting temperature for each of the previously ARD fitted AlSi10 ACE models is determined. This was done using Nested Sampling (NS) which approximates the partition function of an atomic system by exploring the potential energy surface over decreasing energy (or enthalpy) levels, in turn determining the cumulative density of states [61]. From the partition function any thermodynamic quantity can be derived such as the heat capacity. The heat capacity exhibits a signature peak for first-order phase transitions, which includes the liquid-solid transition occurring at the melting temperature. Extensive previous work has shown that NS is an accurate and reliable method for determining the melting temperature [62, 63]. As it explores the entirety of configurational space including gas, liquid and solid phases NS also serves as an excellent test for model robustness. This robustness is partly achieved by the addition of the auxiliary pair potential previously described as V_2 which is added in order to ensure close range repulsion between atoms.

The NS simulations were carried using 896 walkers formed of 32 atom cells (29 Al and 3 Si) using the PYMATNEST software [64]. Starting from the gas phase (initial cell volume of $500 \text{ \AA}^3/\text{atom}$) the walkers explored the potential energy surface by iteratively cloning and decorrelating walkers at decreasing enthalpy levels, passing through the liquid phase and ending up at the ground structure. The decorrelation was done by running MD for 6 timesteps using a 0.1 fs timestep and a total of 1024 MC steps by changing the cell volume, shearing/stretching the cell and swapping atoms (ratio 6:6:6:6) were

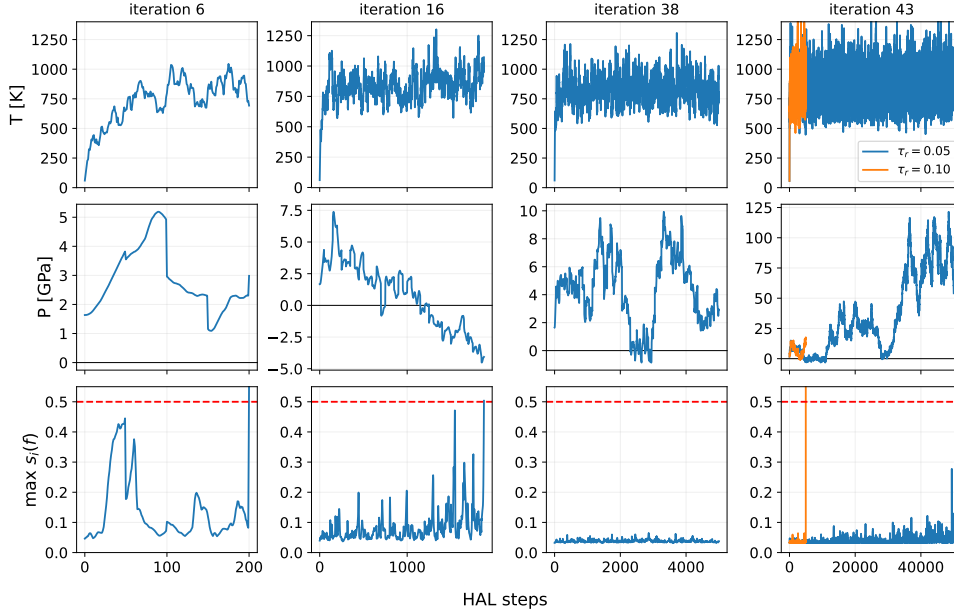


FIG. 3: HAL dynamics for several iterations for the AlSi10 random alloy showing softmax normalised relative force error estimate $s_i(f)$, temperature and pressure. DFT calculations are triggered if the tolerance in red is reached. Pressure fluctuations are due to swap/volume MC steps on HAL potential energy surface E_{HAL} .

performed. The NS cell pressure was set to 0.1 GPa and the minimum aspect ratio of the cell set to 0.85. By summing the enthalpies from the NS simulation the constant pressure partition function is determined from which the heat capacity can be derived through postprocessing. Three independent runs for the each of the ARD models fitted to the AlSi10 HAL database were performed and shown in Fig. 5. All models predicted the expected FCC ground structure, but a difference in predicted melting temperature for varying α' can be seen. Only the $\alpha' = 300k$ and $\alpha' = 80k$ models accurately determine the melting temperature of 867K as predicted by Thermo-Calc with the TCAL4 database [65]. Comparing to Table II demonstrates the relationship between the train/test set error and melting temperature estimates

and suggests that around test set accuracy of 5 meV/atom is required in order to determine the melting temperature accurately.

C. Polyethylene glycol (PEG)

This section presents the application of HAL to build databases for polymers. Polyethylene glycol (PEG) has the formula $\text{H}[\text{OCH}_2\text{CH}_2]_n\text{OH}$, where n is the number of monomer units[66]. From a modelling perspective these polymers are challenging to simulate in vacuum as they form configurations ranging from tightly coiled up to fully stretched out structures. Due to the OH group at the end the polymer can also exhibit hydrogen bonding, which further complicates

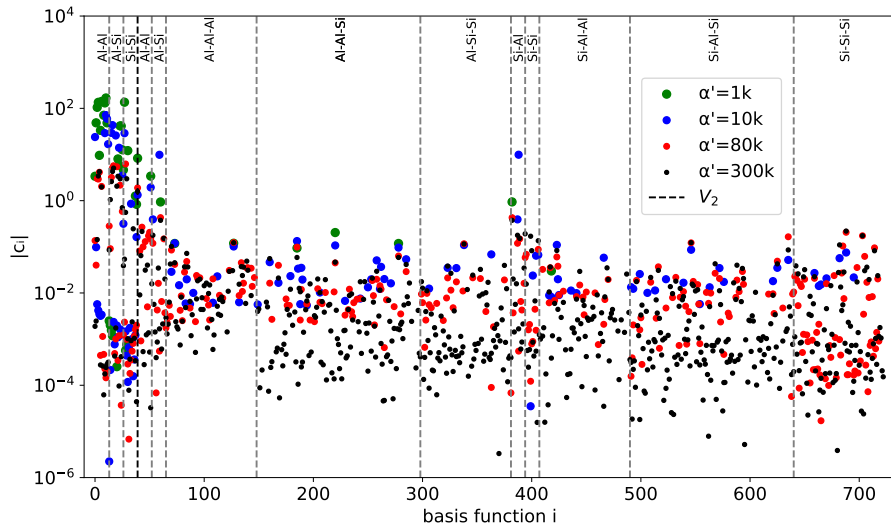


FIG. 4: Coefficient magnitude $|c_i|$ for the 723 basis functions grouped per correlation order and element interaction for various ARD tolerances α' . Large coefficients are assigned to pair interactions, partly captured by the auxiliary pair potential V_2 , as most of the binding energy is contained in these interactions.

its description. These hydrogen bonds typically correspond to low energy configurations and are frequently formed and broken during long MD simulations. This section first presents a benchmark of HAL against AL followed by a demonstration HAL finding configurations exhibiting large errors. Finally, the potential fitted to small polymer units in vacuum is used to predict the density of a long PEG($n=200$) polymer in bulk with excellent accuracy relative to experiment. All DFT reference calculations in this section are carried out with the ORCA code [67] using the ω B97X DFT exchange correlation functional[68] and 6-31G(d) basis set.

1. PEG($n=2$): HAL vs AL

In order to test whether HAL accelerates training database assembly relative to standard AL, a benchmark test was performed. An initial database containing 20 PEG($n=2$) poly-

mers was created by running 500 K NVT molecular dynamics simulation using the general purpose ANI-2x forcefield [11] sampling every 7 ps. This database was fitted using an ACE basis containing basis functions up to correlation order $\nu=3$ and polynomial degree 10 with an outer cutoff 4.5 Å and inner cutoff 0.5 Å. The auxiliary pair potential basis up to polynomial degree 10 and outer cutoff 5.5 Å and did not have an inner cutoff. The weights for the energy w_E , forces w_F were set to 15.0 and 1.0 and remain constant throughout this section on PEG. AL (non-biasing, or $\tau=0.0$) and HAL simulations with varying biasing strengths τ_r were performed using a timestep of 0.5 fs at 500K. Configurations were evaluated using ORCA DFT once $s^{\text{tol}}=0.5$ was reached.

The linear ACE models generated during the AL/HAL simulations were saved and subsequently used in a regular MD stability test and ran for 1 million MD steps at 500 K using a 1 fs timestep for 100 separate runs. A MD simulation was deemed stable if the CC and CO

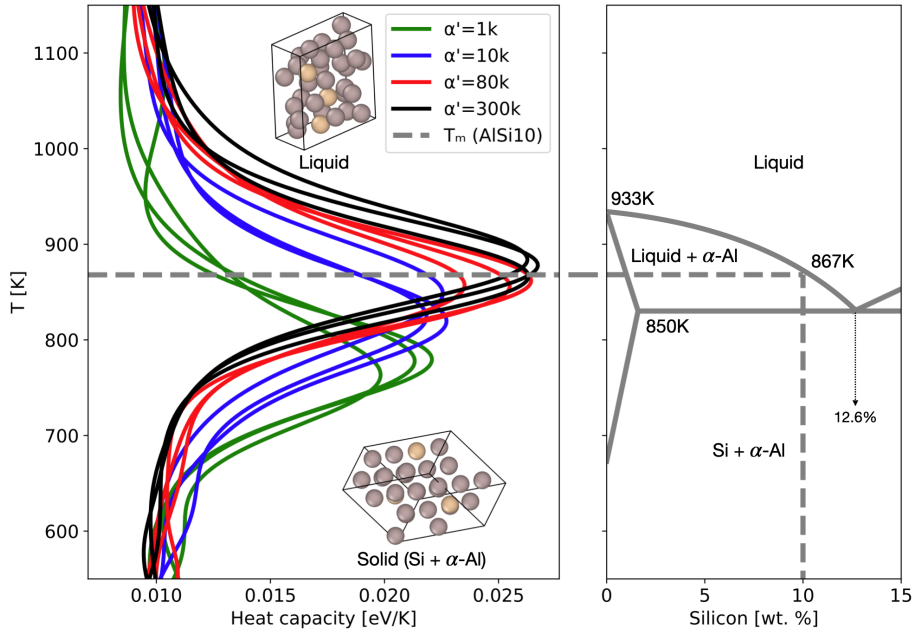


FIG. 5: NS determined heat capacity for ARD fitted linear AlSi10 ACE models (left) and schematic phase diagram for AlSi10 [57] (right). Excellent agreement with the melting temperature is demonstrated for fits with large α' .

bonds along the chain where within 1.0-2.0Å and the CH and OH bonds within 0.8-2.0Å during the simulation. The *minimum* number of stable MD timesteps out of the 100 different simulations is shown in Fig.6 and demonstrates that up to $\tau_r=0.20$ a total of 80 (H)AL iterations are required in order to achieve a minimum MD stability of 1 million steps. The large biasing strength of $\tau_r=0.25$ results in unstable MD dynamics as too strong biasing causes the generation of exceedingly high energy configuration far away from the desired potential energy surface to be included in the training database. Fitting to these configurations leads to a poorly performing model as many unphysical configurations enter the training database resulting.

The HAL run using a biasing strength of $\tau_r=0.20$, achieves minimum 1 million step MD stability after an order of magnitude fewer exploratory MD timesteps compared to standard AL.

2. PEG($n=4$): rare events

Using PEG($n=4$) polymers this section will investigate the ability of HAL to generate and detect configurations with large errors. First a training database was built using the general purpose ANI-2x forcefield [11] at 500K and 800K using a timestep of 1 fs. Configurations were sampled every 7000 timesteps (7 ps), and used to assemble 500K and 800K databases. The 500K database was divided into 750 train configurations and 250 test configurations. The 800K training and test databases both contained 250 configurations. The linear ACE model was extended to include basis functions up to 12 for both the ACE and pair potential, while keeping the cutoffs and correlation order the same ($\nu=3$) too compared to the previous section on PEG($n=2$).

Using the 500K MD sampled training database HAL was started using $\tau_r=0.10$ and a

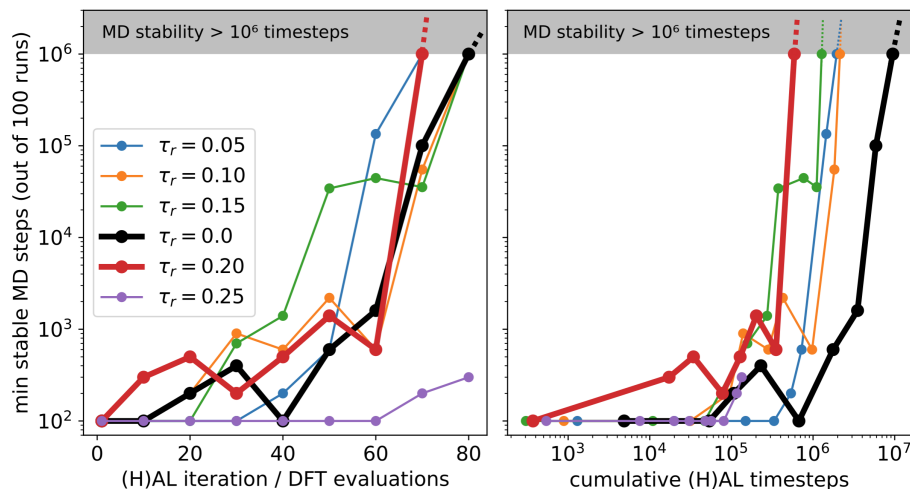


FIG. 6: HAL vs AL benchmark comparing MD stability for 1 million MD steps over 100 seeds. Turning on biasing (non-zero τ_r) creates ACE models achieving stable 100 million MD timestep faster than standard AL by up to an order of magnitude.

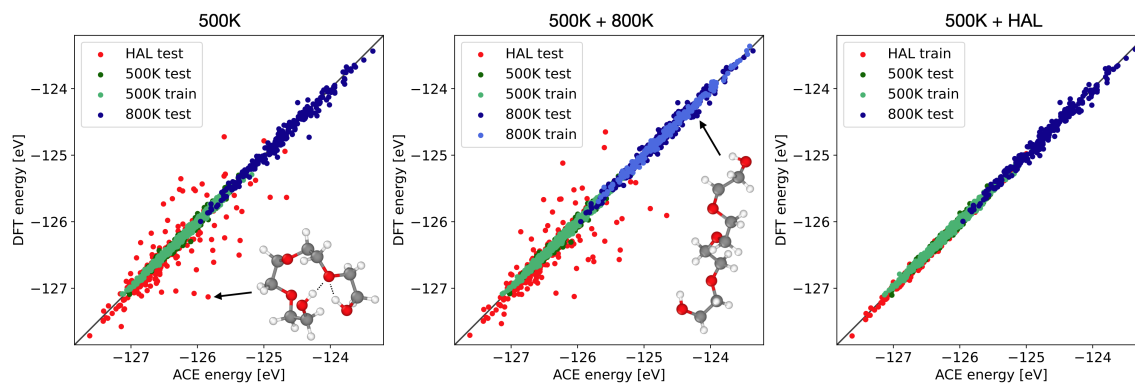


FIG. 7: Energy scatter plots for the 500K (left), 500K+800K (middle) and 500K+HAL (right) ACE models. HAL configuration mostly exhibit (double) hydrogen bonding, or rare events, not contained in the MD 500K/800K decorrelated samples.

timestep of 0.5 fs. The stopping criterion s^{tol} set to 0.5. A total of 200 HAL configurations were generated and formed a HAL database used for both a train and test set. Using the previously described basis three models were created fitted to: 500K, 500K+800K and a 500K+HAL. Energy scatter plots for these three models are shown in Fig. 7 demonstrating that the errors on the HAL-found configurations are large for

both the 500K and 500K+800K fits, despite the fact that these HAL-found configurations are also low in energy! Only by including the HAL configurations in the training database can the errors on these configurations be reduced as shown in Table. III. Inspection of the HAL generated structures exposes a shared characteristic: most of them contain (double) hydrogen bonding across the polymer an example of

	No. configs	500K		500K+800K		500K+HAL	
		E	F	E	F	E	F
500K train	750	30.2	58.3	32.9	60.8	32.4	59.6
500K test	250	49.2	79.3	48.8	76.7	41.6	71.0
800K train	250	-	-	40.0	76.4	-	-
800K test	250	72.7	187.2	67.6	107.7	67.9	102.6
HAL	200	310.9*	427.2*	311.9*	404.6*	47.8 [†]	63.4 [†]

TABLE III: Train and test errors for energies (E) in meV and forces (F) in meV/Å for the 500K, 500K+800K and 500K+HAL databases using ACE. [†] is train error. * is test error.

which is shown in Fig. 7. Such hydrogen-bond formation is a rare event in this system, because only the two ends of the molecule are capable of hydrogen bonding. It is difficult to find these configurations using regular MD (even when using elevated temperatures), whereas HAL finds them easily.

3. Bulk PEG($n=200$) density

As a final investigation the density of a PEG($n=200$) polymer containing 1400 atoms is determined using an ACE model fitted to a HAL generated PEG training database containing polymer sizes ranging from $n=2$ to $n=32$ monomer units. This database contained configurations from the previous PEG sections and extended using configurations sized $n = 8$, $n = 16$ and $n = 32$. The training database included *standard* ANI MD sampled configurations at 500K including 1000 PEG($n=4$) configurations (from the previous section), as well as 50 PEG($n=2$), 100 PEG($n=8$), 100 PEG($n=16$) and 18 PEG($n=32$) configurations. Starting from this data HAL was used to generate an extra 64 PEG($n=16$) and 91 PEG($n=32$) HAL configurations until dynamics was deemed stable. The linear ACE basis used for the regression task was identical to the ACE in the previous section on PEG($n=4$), and any force components with greater than 20 eV/Å were excluded from the fit.

Using the ACE model a PEG($n=200$) polymer was simulated in LAMMPS [70] with the PACE evaluator pair style with periodic bound-

ary conditions. Since the training database only contained small polymers segments in vacuum this periodic simulation demonstrates a large degree of extrapolation to configurations far away from the training database. Furthermore, the DFT code used to evaluate the training energies and forces does not support periodic boundary conditions making DFT simulation of the 1400 atom PEG($n=200$) simulation box not just computationally infeasible, but practically impossible in this case.

The resulting linear ACE model was timed at 220 μ s/atom/core per MD step. LAMMPS NPT simulations were performed at 1 bar using a 1 fs timestep at 300K, 400K, 500K and 600K. The recorded density as a function of simulation time is plotted in Fig. 8. Using the last 500 ps from the 300K simulation the density was determined to be 1.238 g/cm³. This value is around 3% higher than the experimental value of 1.2 g/cm³ [69].

III. METHODS

A. Hyperactive Learning (HAL)

The HAL potential energy E_{HAL} as defined in (1) biases MD simulations during the exploration step in AL towards uncertainty by shifting the potential energy surface and assigning lower energies to configurations with high uncertainty. We have shown that when σ is the standard deviation of the posterior-predictive uncertainty of Energy, it can be computationally cheaply approximated by a Monte-Carlo es-

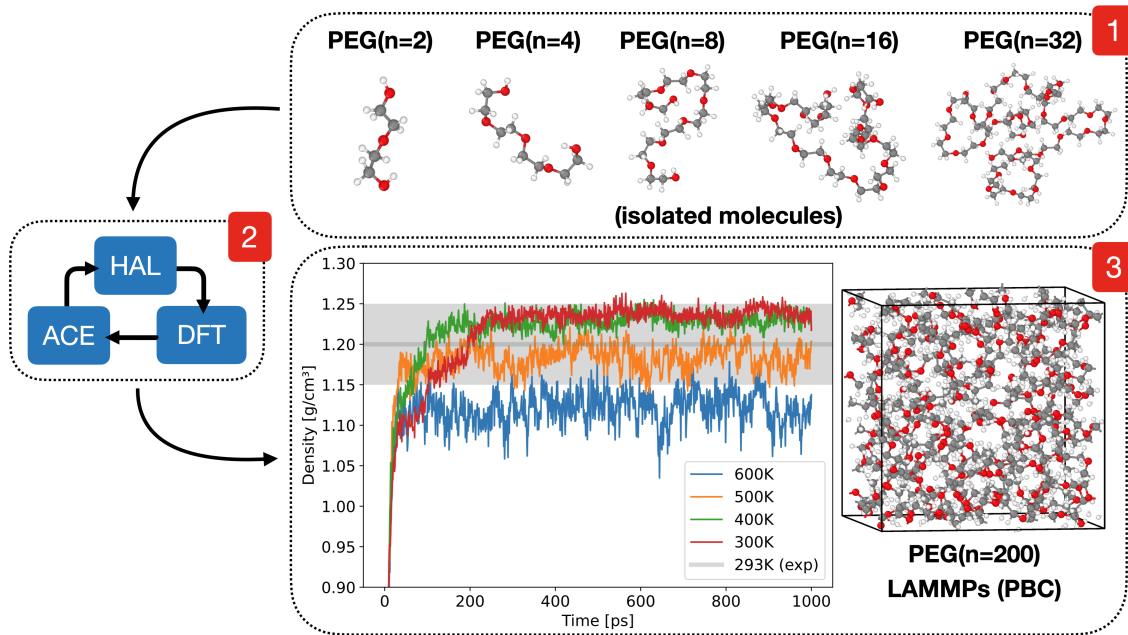


FIG. 8: HAL protocol for building linear ACE PEG model accurately determining PEG($n=200$) density within experimental accuracy of 1.2 g/cm^3 at 297K (shaded area) [69]. Training database only included small polymers ranging from $n=2$ to $n=32$ in isolation.

estimate $\tilde{\sigma}$; see (5). Likewise, the derivative of $\tilde{\sigma}$ can be computed as

$$\nabla \tilde{\sigma} = \frac{\nabla \tilde{\sigma}^2}{2\tilde{\sigma}} \quad (9)$$

where

$$\begin{aligned} \nabla \tilde{\sigma}^2 &= \frac{2}{K} \sum_{k=1}^K (E^k - \bar{E}) (\nabla E^k - \nabla \bar{E}) \\ &= \frac{2}{K} \sum_{k=1}^K (E^k - \bar{E}) (\bar{F} - F^k) \end{aligned} \quad (10)$$

and $F^k = -\nabla E^k$, $\bar{F} = -\nabla \bar{E}$. These predictions are obtained by ensemble parameterisations $\{\mathbf{c}_k\}_{k=1}^K$, while $\bar{\mathbf{c}}$ is the analytic mean of the posterior distribution as specified in (22). The sum over K is over the ensemble or committee of models, which in this work was chosen to be linear (ACE) models. Other architectures such as neural networks ensembles may be considered in future work. This quantity in essence

is a computationally cheap method of determining the gradient towards (total) energy uncertainty and may be interpreted as a conservative *biasing force*,

$$F^{\tilde{\sigma}} := \nabla \tilde{\sigma}. \quad (11)$$

HAL dynamics adds this biasing force to MD in order to accelerate the generation of configurations with high uncertainty, which sets HAL apart from AL. Setting $\tau=0$ recovers standard MD dynamics, and in this sense, HAL generalizes AL. Interestingly, previous work employed a biasing force using a neural network interatomic potential [71] but biased *away* from uncertainty in order to stabilise the MD dynamics.

The biasing strength τ can either be set as a constant or adapted during the HAL simulation. Controlling the biasing strength is important as too strong biasing can quickly lead to unphysical configurations, whereas low biasing generates valuable configurations at a slow rate. The

adaptive biasing works by first setting τ_r and performing a burn-in period to record the magnitudes (or, norms) of $F^{\tilde{\sigma}}$ and \bar{F} . Typically, the burn-in period is set to the latest 100 timesteps δt . The biasing strength τ is then determined by satisfying

$$\tau_r = \frac{\tau \sum_{m=1}^{100} \|\bar{F}(t - m\Delta t)\|}{\sum_{m=1}^{100} \|F^{\tilde{\sigma}}(t - m\Delta t)\|}. \quad (12)$$

The new parameter τ_r is generally set between 0.05 and 0.25. It can be understood as the approximate relative average strength of the biasing force in comparison to the average force of the fitted model. Using this adaptive biasing term aids usability and tunes the biasing strength to ensure that HAL gently drives MD towards high uncertainty. The value may loosely be interpreted as the relative magnitude of the biasing force compared to the true gradient of the potential energy surface. Larger τ_r increases the biasing strength and rate at which configurations with high uncertainty are generated. In order to sample configurations at desired pressures and temperatures a proportional control barostat was added as well as a Langevin thermostat.

B. Atomic Cluster Expansion (ACE)

The ACE model decomposes the total energy E of a configuration R as a sum of parameterised atomic energies,

$$E(\mathbf{c}; R) = \sum_{i \in R} E_i(\mathbf{c}; R). \quad (13)$$

The atomic energies E_i are then parameterised by a linear model, $E_i = \mathbf{c} \cdot \mathbf{B}_i$, where \mathbf{B}_i denotes the ACE basis. The present work employs a particularly simple variant, which we review briefly: Given relative atomic positions $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$ and associated chemical elements one evaluates a one-particle basis

$$\phi_{znlm}(\mathbf{r}_{ji}, z_j) = \delta_{zz_j} R_n(r_{ji}) Y_{lm}(\hat{\mathbf{r}}_{ji}), \quad (14)$$

followed by a pooling operation resulting in features

$$A_{iznml} = \sum_j \phi_{znlm}(\mathbf{r}_{ji}, z_j), \quad (15)$$

that are denoted the *atomic basis* in the context of the ACE model. Taking a ν order (tensor) product results in many-body correlation functions incorporating $(\nu+1)$ body-order interactions,

$$\mathbf{A}_{iznlm} = \prod_{t=1}^{\nu} A_{iz_t n_t l_t m_t}. \quad (16)$$

The \mathbf{A} -basis is a complete basis of permutation-invariant functions but does not incorporate rotation or reflection symmetry. An isometry invariant basis \mathbf{B} is constructed by averaging over rotations and reflections. Representation theory of the orthogonal group $O(3)$ shows that this can be expressed as a sparse linear operation and results in

$$\mathbf{B}_i = \mathbf{C} \mathbf{A}_i, \quad (17)$$

where \mathbf{C} contains generalised Clebsch-Gordan coefficients; we refer to [24, 25] for further details.

A major benefit of the linear ACE model is that the computational cost of evaluating a site energy E_i scales only linearly with the number of neighbouring atoms, as well as with the body order $\nu + 1$.

C. (Bayesian) Linear Regression

The parameters of linear ACE models are fitted by solving a linear regression problem. The associated squared loss function $L(\mathbf{c})$ to be minimised over configurations R in training set \mathbf{R} with corresponding (DFT) observations for energy \mathcal{E}_R , forces \mathcal{F}_R is

$$L(\mathbf{c}) = \sum_{R \in \mathbf{R}} (w_E |E(\mathbf{c}; R) - \mathcal{E}_R|^2 + w_F |F(\mathbf{c}; R) - \mathcal{F}_R|^2) \quad (18)$$

where w_E and w_F are weights specifying the relative importance of the DFT observations. When fitting materials a third term is added $w_V|V(\mathbf{c}; R) - \mathcal{V}_R|^2$ referring to the virial stress components of the configuration R . This minimisation problem can be recast in the standard form

$$\arg \min_{\mathbf{c}} \|\mathbf{y} - \Psi \mathbf{c}\|^2 + \eta \|\mathbf{c}\|^2, \quad (19)$$

where Ψ is the design matrix and \mathbf{y} collects the observations to which the parameters are fitted. Here, we also added a Tychonov regularisation with regularisation parameter $\eta > 0$ which is commonly determined through a model selection criterion such as cross-validation.

This linear regression model can be cast in a Bayesian framework by specifying a prior distribution $p(\mathbf{c})$ over the regression parameters, and an (additive) probabilistic error model ϵ , which gives rise to the generative model

$$\mathbf{y} = \Psi \mathbf{c} + \epsilon. \quad (20)$$

In the context of this work ϵ models random perturbations of DFT calculations and is assumed to be mainly present due to the locality assumption and DFT convergence properties, e.g. k-point sampling. For simplicity the noise ϵ is in this work assumed to be statistically independent across observations and Gaussian distributed with zero mean and precision (inverse variance) λ , but in principle extensions to other noise models can be made. This noise model gives rise to the likelihood function

$$\begin{aligned} p(\mathbf{y}|R, \mathbf{c}, \lambda) &= \prod_{i=1}^{N_{\text{obs}}} \mathcal{N}(\mathbf{y}_i | M(\mathbf{c}; R), \lambda^{-1}) \\ &= \left(\frac{\lambda}{2\pi}\right)^{N_{\text{obs}}/2} \exp\left\{-\frac{\lambda}{2}\|\mathbf{y} - \Psi \mathbf{c}\|^2\right\} \end{aligned} \quad (21)$$

where $M(\mathbf{c}; R)$ refers to the corresponding regression function for energy, force or virial.

By restricting ourselves to a Gaussian error model, and assuming the prior to be Gaussian as well, i.e., $p(\mathbf{c}) = \mathcal{N}(\mathbf{c}; \mathbf{0}, \Sigma_0)$, it is ensured that the posterior distribution, $\pi(\mathbf{c}) =$

$p(\mathbf{c}|\mathbf{R}, \mathbf{y}, \Sigma_0, \lambda)$, is Gaussian with closed form expressions for both the distribution mean $\bar{\mathbf{c}}$ and variance Σ ,

$$\begin{aligned} \bar{\mathbf{c}} &= \lambda \Sigma \Psi^T \mathbf{y} \\ \Sigma^{-1} &= \Sigma_0^{-1} + \lambda \Psi^T \Psi. \end{aligned} \quad (22)$$

In the context of this work, having closed form expressions for both these quantities is desirable as it (i) allows for conceptual easy and fast generation of independent samples $\{\mathbf{c}^k\}_{k=1}^K$ from the posterior distribution, and (ii) allows for a parametrisation of the fitted model with the exact mean, $\bar{\mathbf{c}}$, of the posterior distribution.

In what follows we briefly describe two Bayesian regression techniques, Bayesian Ridge Regression (BRR), which we use to produce Bayesian fits during the HAL data generation phase, and the more expensive Automatic Relevance Determination (ARD), which we use to obtain a final model fit after the data generation is complete.

D. Bayesian Ridge Regression (BRR)

In Bayesian Ridge Regression the covariance of the prior is assumed to be isotropic, i.e.,

$$p(\mathbf{c}|\alpha) = \mathcal{N}(\mathbf{c}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \quad (23)$$

for some hyper-parameter $\alpha > 0$, the precision of the prior distribution.

Under this choice of prior, the logarithm of the posterior distribution takes the form

$$\ln \pi(\mathbf{c}) = -\frac{\lambda}{2}\|\mathbf{y} - \Psi \mathbf{c}\|^2 - \frac{\alpha}{2}\|\mathbf{c}\|^2 + C, \quad (24)$$

where C is some constant. Thus, maximising the (log-)posterior for this choice of prior, is equivalent to solving the regularised least square problem Eq. 24 with ridge penalty $\eta = \alpha/\lambda$. This shows that the prior naturally gives rise to a regularised solution, keeping coefficient parameters small.

The determination of the hyper-parameters α and λ in BRR is achieved by optimising the marginal log likelihood also known as evidence

maximisation [72]. One first defines the evidence function as

$$p(\mathbf{y}|\alpha, \lambda) = \int p(\mathbf{y}|\mathbf{c}, \beta)p(\mathbf{c}|\alpha)d\mathbf{c} \quad (25)$$

which marginalises out the coefficients \mathbf{c} and describes the likelihood of observing the data given the hyperparameters α and λ . Using the previously defined definitions the evidence function can be expressed as

$$p(\mathbf{y}|\alpha, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{N_{\text{obs}}/2} \left(\frac{\alpha}{2\pi}\right)^{N_{\text{basis}}/2} \int \exp\left\{-\frac{\lambda}{2}\|\mathbf{y} - \Psi\mathbf{c}\|^2 - \frac{\alpha}{2}\|\mathbf{c}\|^2\right\} \quad (26)$$

where N_{basis} is the dimensionality of \mathbf{c} . Completing the square in the exponent and taking the log gives rise to the marginal log likelihood

$$\begin{aligned} \ln p(\mathbf{y}|\alpha, \lambda) &= \frac{N_{\text{basis}}}{2} \ln \alpha + \frac{N_{\text{obs}}}{2} \ln \lambda \\ &\quad - \frac{\lambda}{2}\|\mathbf{y} - \Psi\mathbf{c}\|^2 - \frac{\alpha}{2}\|\mathbf{c}\|^2 + \\ &\quad \frac{1}{2} \ln \|\Sigma\| - \frac{N}{2} \ln(2\pi) \end{aligned} \quad (27)$$

which can be maximised with respect to α and λ in order to maximise the marginal likelihood and obtain the statistically most probable solution given the basis and data.

E. Automatic Relevance Determination (ARD)

Automatic Relevance Determination (ARD) modifies BRR by relaxing the isotropy of the prior and assigning a hyperparameter α_i to independently regularise each coefficient c_i . The corresponding prior is given by

$$\begin{aligned} p(\mathbf{c}|\alpha) &= \mathcal{N}(\mathbf{c}|\mathbf{0}, \mathcal{A}^{-1}) \\ \mathcal{A} &= \text{diag}(\alpha_1, \dots, \alpha_{N_{\text{basis}}}). \end{aligned} \quad (28)$$

This prior determines the relevance of each parameter c_i , or basis function, which effectively results in a feature selection. Basis functions are

ranked based on their relevance and are pruned if determined irrelevant, in turn producing a sparse solution. In practice, sparse models obtained through ARD often yield better generalisation than BRR. Using ARD requires the specification of a threshold parameter α' setting the minimum relevance of basis functions included in the fit. Adjusting this parameter controls the balance between accuracy and sparsity of the model.

F. Posterior Predictive Distribution

A key property of the Bayesian approach is that it provides a way to quantify uncertainty in terms of the posterior-predictive distribution, which accounts both for parameter uncertainty as given by the posterior distribution as well as uncertainty due to observation error. For the observations we consider (energy, forces, and virials) predictions y of these quantities at a atomic configuration R can probabilistically be described as

$$\begin{aligned} y &= M(\mathbf{c}; R) + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \lambda\mathbf{I}_d), \\ \mathbf{c} &\sim \pi(\mathbf{c}). \end{aligned} \quad (29)$$

where $M(\mathbf{c}, R)$ denotes the respective regression model and d is the dimension of the predicted quantity (e.g. $d=9$, for predictions of virials). The distribution of the prediction y , denoted as $p(y|R)$, is referred to as the posterior-predictive distribution of y and can be verified to have the generic form

$$p(y|R) = \int p(y|R, \mathbf{c})\pi(\mathbf{c})d\mathbf{c}. \quad (30)$$

where under the here assumed noise model

$$p(y|R, \mathbf{c}) = \mathcal{N}(y|M(\mathbf{c}, R), \mathbf{I}_d\lambda^{-1}). \quad (31)$$

In particular, the integral in Eq. (30) collapses to the Gaussian density for predictions of energy, i.e.,

$$p(E|R) = \mathcal{N}(E|\bar{\mathbf{c}} \cdot \mathbf{B}, \sigma_E^2), \quad (32)$$

where the variance σ_E^2 is as specified in (3).

IV. DATA AVAILABILITY

The data will be made available at time of publication.

V. CODE AVAILABILITY

The HAL code will be made available at time of publication.

VI. ACKNOWLEDGEMENTS

GC and CvdO acknowledge the support of UKCP grant number EP/K014560/1. CvdO would like to acknowledge the support of EPSRC (Project Reference: 1971218) and Dassault Systèmes UK. CO acknowledges support

of the NSERC Discovery Grant (IDGR019381) and the NFRF Exploration Grant GR022937.

VII. AUTHOR CONTRIBUTIONS

CvdO developed efficient uncertainties for ACE models. MS conceived the idea of uncertainty-biasing of AL exploration. CvdO led the implementation of the HAL framework. DK helped generate data. CvdO wrote the first version of the manuscript. All authors discussed the theory and results and edited the manuscript.

VIII. COMPETING INTERESTS

The authors declare no competing interests.

-
- [1] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.
- [2] Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141, 08 2021.
- [3] John A. Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical Reviews*, 121(16):9816–9872, 08 2021.
- [4] A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics*, 285:316–330, 2015.
- [5] Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E. Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228:178–184, 2018.
- [6] Ivan S Novikov, Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. The MLIP package: moment tensor potentials with MPI and active learning. *Mach. Learn.: Sci. Technol.*, 2(2):025002, December 2020.
- [7] Gabriele C. Sosso, Volker L. Deringer, Stephen R. Elliott, and Gábor Csányi. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Molecular Simulation*, 44(11):866–880, 2018.
- [8] Volker L. Deringer, Noam Bernstein, Gábor Csányi, Chiheb Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A. Drabold, and Stephen R. Elliott. Origins of structural and electronic transitions in disordered silicon. *Nature*, 589(7840):59–64, 2021.
- [9] Venkat Kapil, Christoph Schran, Andrea Zen, Ji Chen, Chris J. Pickard, and Angelos Michaelides. The first-principles phase diagram of monolayer nanoconfined water. *Nature*, 609(7927):512–516, 2022.
- [10] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007.

- [11] Christian Devereux, Justin S. Smith, Kate K. Huddleston, Kipton Barros, Roman Zubatyuk, Olexandr Isayev, and Adrian E. Roitberg. Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *Journal of Chemical Theory and Computation*, 16(7):4192–4202, 2020. PMID: 32543858.
- [12] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203, 2017.
- [13] Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228:178–184, 2018.
- [14] Albert P Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015.
- [15] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12):1–9, 2017.
- [16] Anders S. Christensen, Lars A. Bratholm, Felix A. Faber, and O. Anatole Von Lilienfeld. FCHL revisited: Faster and more accurate quantum machine learning. *Journal of Chemical Physics*, 152(4), 2020.
- [17] Stefan Chmiela, Huziel E. Sauceda, Igor Poltavsky, Klaus Robert Müller, and Alexandre Tkatchenko. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45, 2019.
- [18] Bastiaan J. Braams and Joel M. Bowman. Permutationally invariant potential energy surfaces in high dimensionality. *International Reviews in Physical Chemistry*, 28(4):577–606, 2009.
- [19] Cas van der Oord, Geneviève Dusson, Gábor Csányi, and Christoph Ortner. Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials. *Machine Learning: Science and Technology*, 1(1):015004, 2020.
- [20] Alice E A Allen, Geneviève Dusson, Christoph Ortner, and Gábor Csányi. Atomic permutationally invariant polynomials for fitting molecular force fields. *Machine Learning: Science and Technology*, 2(2):025017, 2021.
- [21] Aidan P Thompson, Laura P Swiler, Christian R Trott, Stephen M Foiles, and Garritt J Tucker. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics*, 285:316–330, 2015.
- [22] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010.
- [23] Alexander V. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.
- [24] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.
- [25] Geneviève Dusson, Markus Bachmayr, Gábor Csányi, Ralf Drautz, Simon Etter, Cas van der Oord, and Christoph Ortner. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946, 2022.
- [26] Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice EA Allen, Daniel J Cole, Christoph Ortner, and Gábor Csányi. Linear atomic cluster expansion force fields for organic molecules: beyond rmse. *Journal of chemical theory and computation*, 17(12):7696–7711, 2021.
- [27] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:6–13, 2017.
- [28] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019.
- [29] Oliver T. Unke and Markus Meuwly. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation*, 15(6):3678–3693, 2019.
- [30] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional Message Passing for Molecular Graphs. *arXiv:2003.0312*, pages 1–13, 2020.
- [31] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. pages 9377–9388, 2021.

- [32] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.
- [33] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022.
- [34] Ilyes Batatia, Dávid Péter Kovács, Gregor NC Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *arXiv preprint arXiv:2206.07697*, 2022.
- [35] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), 2017.
- [36] K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K. R. Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 2017-Decem(1):992–1002, 2017.
- [37] Patrick Rowe, Volker L. Deringer, Piero Gasparotto, Gábor Csányi, and Angelos Michaelides. An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics*, 153(3):034702, 2020.
- [38] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X*, 8:041048, Dec 2018.
- [39] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature Communications*, 11(1):5461, 2020.
- [40] Evgeny V. Podryabinkin and Alexander V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.
- [41] Konstantin Gubaev, Evgeny V. Podryabinkin, Gus L.W. Hart, and Alexander V. Shapeev. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Computational Materials Science*, 156:148–156, 2019.
- [42] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6:1–11, 2020.
- [43] Jonathan Vandermause, Yu Xie, Jin Soo Lim, Cameron J. Owen, and Boris Kozinsky. Active learning of reactive bayesian force fields applied to heterogeneous catalysis dynamics of h/pt. *Nature Communications*, 13(1):5183, 2022.
- [44] Ganesh Sivaraman, Anand Narayanan Krishnamoorthy, Matthias Baur, Christian Holm, Marius Stan, Gábor Csányi, Chris Benmore, and Álvaro Vázquez-Mayagoitia. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials*, 6(1):104, 2020.
- [45] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.
- [46] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Physical review letters*, 96(9):090601, 2006.
- [47] Simone Marsili, Alessandro Barducci, Riccardo Chelli, Piero Procacci, and Vincenzo Schettino. Self-healing umbrella sampling: a nonequilibrium approach for quantitative free energy calculations. *The Journal of Physical Chemistry B*, 110(29):14011–14013, 2006.
- [48] Bradley M Dickson, Frédéric Legoll, Tony Lelièvre, Gabriel Stoltz, and Paul Fleurat-Lessard. Free energy calculations: An efficient adaptive biasing potential method. *The Journal of Physical Chemistry B*, 114(17):5823–5830, 2010.
- [49] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of chemical physics*, 115(20):9169–9183, 2001.
- [50] Jérôme Hénin and Christophe Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *The Journal of chemical physics*, 121(7):2904–2914, 2004.
- [51] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

- [52] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.
- [53] James T. Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for bayesian optimization, 2018.
- [54] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.
- [55] Ron Wehrens, Hein Putter, and Lutgarde M.C Buydens. The bootstrap: a tutorial. *Chemo-metrics and Intelligent Laboratory Systems*, 54(1):35–52, 2000.
- [56] Nicholas Lubbers et al. Maksim Kulichenko, Kipton Barros. Uncertainty driven dynamics for active learning of interatomic potentials. <https://doi.org/10.21203/rs.3.rs-2109927/v1>, *Research Square (PREPRINT)*, 2022.
- [57] F. Alghamdi and M. Haghshenas. Microstructural and small-scale characterization of additive manufactured als10mg alloy. *SN Applied Sciences*, 1(3):255, 2019.
- [58] Stewart J. Clark, Matthew D. Segall, Chris J. Pickard, Phil J. Hasnip, Matt I. J. Probert, Keith Refson, and Mike C. Payne. First principles methods using castep. *Zeitschrift für Kristallographie - Crystalline Materials*, 220(5-6):567–570, 2005.
- [59] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.
- [60] Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, et al. Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12, 2021.
- [61] Livia B. Pártay, Gábor Csányi, and Noam Bernstein. Nested sampling for materials. *The European Physical Journal B*, 94(8):159, 2021.
- [62] Robert JN Baldock, Livia B Pártay, Albert P Bartók, Michael C Payne, and Gábor Csányi. Determining pressure-temperature phase diagrams of materials. *Physical Review B*, 93(17):174108, 2016.
- [63] Livia B Pártay. On the performance of interatomic potential models of iron: Comparison of the phase diagrams. *Computational Materials Science*, 149:153–157, 2018.
- [64] N. Bernstein R. J. N. Baldock L. B. Partay J. R. Kermode T.D. Daff A.P. Bartok and G. Csanyi. <https://github.com/libatoms/pymatnest>. 2016.
- [65] Ming Tang, P. Chris Pistorius, Sneha Narra, and Jack L. Beuth. Rapid solidification: Selective laser melting of als10mg. *JOM*, 68(3):960–966, 2016.
- [66] Z. Karimi, L. Karimi, and H. Shokrollahi. Nano-magnetic particles used in biomedicine: Core and coating materials. *Materials Science and Engineering: C*, 33(5):2465–2475, 2013.
- [67] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The orca quantum chemistry program package. *The Journal of Chemical Physics*, 152(22):224108, 2020.
- [68] Jeng-Da Chai and Martin Head-Gordon. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of Chemical Physics*, 128(8):084106, 2008.
- [69] *Polyethylene glycol [MAK Value Documentation, 1998]*, pages 248–270. John Wiley and Sons, Ltd, 2012.
- [70] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022.
- [71] Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153(10):104105, 2020.
- [72] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.