

Hyperactive learning for data-driven interatomic potentials

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

DOI:

[10.1038/s41524-023-01104-6](https://doi.org/10.1038/s41524-023-01104-6)

License:

Creative Commons: Attribution (CC BY)

Document Version

Publisher's PDF, also known as Version of record

Citation for published version (Harvard):

van der Oord, C, Sachs, M, Kovács, DP, Ortner, C & Csányi, G 2023, 'Hyperactive learning for data-driven interatomic potentials', *npj Computational Materials*, vol. 9, no. 1, 168. <https://doi.org/10.1038/s41524-023-01104-6>

[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.

ARTICLE OPEN



Hyperactive learning for data-driven interatomic potentials

Cas van der Oord^{1✉}, Matthias Sachs², Dávid Péter Kovács¹, Christoph Ortner³ and Gábor Csányi¹

Data-driven interatomic potentials have emerged as a powerful tool for approximating ab initio potential energy surfaces. The most time-consuming step in creating these interatomic potentials is typically the generation of a suitable training database. To aid this process hyperactive learning (HAL), an accelerated active learning scheme, is presented as a method for rapid automated training database assembly. HAL adds a biasing term to a physically motivated sampler (e.g. molecular dynamics) driving atomic structures towards uncertainty in turn generating unseen or valuable training configurations. The proposed HAL framework is used to develop atomic cluster expansion (ACE) interatomic potentials for the AlSi10 alloy and polyethylene glycol (PEG) polymer starting from roughly a dozen initial configurations. The HAL generated ACE potentials are shown to be able to determine macroscopic properties, such as melting temperature and density, with close to experimental accuracy.

npj Computational Materials (2023)9:168; <https://doi.org/10.1038/s41524-023-01104-6>

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 centred symmetry functions¹⁰ and have been used in models such as ANI^{11,12} and DeepMD⁵. Another widely used approach is Gaussian process regression (GPR) implemented in models such as SOAP/GAP^{13,14}, FCHL¹⁵ and sGDML¹⁶. Linear approximations of the PES have also been introduced initially by using permutation invariant polynomials (PIPs)¹⁷ and the more recent atomic PIPs variant^{18,19}. Other linear models include spectral neighbour analysis potentials⁴ based on the bispectrum²⁰, moment tensor potentials²¹ and the atomic cluster expansion (ACE)^{22–24}. More recently, message passing neural network (MPNN) architectures have been introduced^{25–31} the most recent of which have been able to outperform any of the previously mentioned models regarding accuracy on benchmarks such as MD17³² and ISO17³³. Central to all of these models is that they are fitted to a training database comprised of configurations R labelled with total energy \mathcal{E}_R , forces \mathcal{F}_R and perhaps virial stress \mathcal{V}_R observations, obtained from electronic structure calculations. 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^{34–36}. 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 numerically unstable 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^{37,38} and the Gaussian process (GP) based methods FLARE^{39,40} and GAP⁴¹ 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 (HAL) 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

¹University of Cambridge, Cambridge CB2 1PZ, UK. ²University of Birmingham, Birmingham B15 2TT, UK. ³University of British Columbia, Vancouver, BC V6T 1Z2, Canada.

✉email: casv2@cam.ac.uk

unreliably predicted configurations but retain the overall focus on low energy configurations carrying large contributions to the partition function. This proposed framework is reminiscent of the exploration-exploitation trade-off originating from Bayesian optimisation (BO), a technique used to efficiently optimise a computationally expensive “black box” function. BO has been shown to yield state-of-the-art results for optimisation problems while simultaneously minimising incurred computational costs by requiring fewer ground-truth evaluations⁴². In the wider community BO is seen as a type of AL, and so is the proposed HAL framework in this work. The novelty of this work is combining MD with BO to accelerate the development of data-driven interatomic potentials.

BO has been applied to atomistic systems previously in global structure search^{43–46} where the PES is optimised to find stable structures. Other previous work balancing exploration and exploitation in data-driven interatomic potentials is also closely related, where configurations were generated by balancing high uncertainty and high-likelihood (or rather low-energy)⁴⁷. Here the PES was explored by perturbing geometries while monitoring uncertainty rather than explicitly running MD. Note that upon the completion of this work, we discovered a closely related work that also uses uncertainty-biased MD⁴⁸. The two studies were performed independently, and appeared on preprint servers near-simultaneously.

In BO, an acquisition function balances exploration and exploitation, controlled by a biasing parameter. In our hyperactive learning framework, the HAL potential energy surface E_{HAL} :

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

takes on a similar role. Here, E is the predicted potential energy and σ is an uncertainty measure, which in this work is set to be the standard deviation of predicted total energy. The parameter τ , referred to as the biasing strength, controls the exploration of unseen parts of the PES and needs to be carefully tuned in order for the HAL-MD trajectory to remain energetically sensible. This is achieved by the introduction of an on-the-fly auto-tuning scheme using a relative biasing parameter τ_r (see § “Methods” for details). The addition of a biasing potential has a long history in the study of rare events and free energy computations, using adaptive biasing strategies such as meta-dynamics^{49,50}, umbrella sampling^{51,52}, and similar methods (e.g. refs. ^{53,54}). 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 towards increasing uncertainty corresponding to regions of configuration space not accurately described by the training data. Viewing HAL as an adaptive-biasing technique also contrasts it against more aggressive AL approaches that explore configuration space via thermostated MD at high temperature. In the latter case all degrees of freedom are indiscriminately accelerated. In the absence of strong energetic barriers this drastically increases the size of the sampled configurational space rendering an exhaustive exploration of physically relevant configurations infeasible. In contrast, HAL only accelerates the degrees of freedom in the direction of increasing uncertainty. Intuitively, one may expect that this keeps the size of sampled configurational space constrained and exploration effective.

Choosing σ to be the predicted energy’s standard deviation makes E_{HAL} coincide exactly with the lower confidence bound (LCB), which is a commonly used acquisition function in BO. In particular, it has previously been used to optimise the potential energy surface⁴³ rather than to sample the corresponding statistical ensembles as performed in this work. From both a theoretical and modelling perspective other versions of HAL are of high interest. For example, we expect that using the relative force uncertainties that we introduce below as biasing potentials, would result in a more targeted biasing that is consistent with the proposed stopping criterion presented in Eq. (6). However, since such a formulation of HAL would require the evaluation of higher order derivatives of the predicted energy, we leave this to future work.

We make the general HAL concept concrete in the context of the ACE “machine learning potential” framework^{22,23}, however, the methods we propose can be directly applied to any linear models and Gaussian process type models, and are in principle also extendable to any other ML potential that comes with an uncertainty measure, including deep neural network models. Different methods of setting up such ensembles or committees exist for linear, GP or NN frameworks, such as dropout⁵⁵, or bootstrapping⁵⁶. In the context of Bayesian models, ensembles can be obtained as Monte Carlo samples from the corresponding posterior distribution. More specifically, considering linear ACE, the site (or atomic) energy is expressed as follows:

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

and the total energy E is defined as $E = \sum_i E_i = \mathbf{c} \cdot \mathbf{B}$ where $\mathbf{B} = \sum_i \mathbf{B}_i$ is the linear ACE basis. Due to this linearity, implementing a Bayesian model formulation is particularly straightforward allowing for efficient and analytical uncertainty estimation as described in § “Bayesian ridge regression (BRR)”. Assuming an isotropic Gaussian prior on the model parameters and Gaussian independent and identically distributed (i.i.d) noise on observations, yields an explicit definition of the standard deviation of the posterior predictive distribution:

$$\sigma = \sqrt{\frac{1}{\lambda} + \mathbf{B}^T \mathbf{\Sigma} \mathbf{B}}, \quad (3)$$

which has energy units, in correspondence with Eq. (1). Here, the covariance matrix $\mathbf{\Sigma}$ is defined as:

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

and α, λ are hyperparameters whose treatment are detailed in the Methods section, and $\mathbf{\Psi}$ is the design matrix of the linear regression problem.

The evaluation of σ in Eq. (3) is computationally expensive for a large basis \mathbf{B} ; scaling as $O(N_{\text{basis}}^2)$. To improve computational efficiency, σ can be approximated by a committee of K potentials parameterised by an ensemble of parameters $\{\mathbf{c}^k\}_{k=1}^K$ that are obtained by sampling from the posterior distribution $\pi(\mathbf{c})$ (see Eq. (25) for further details). This gives rise to K committee energy predictions, $E^k = \mathbf{c}^k \cdot \mathbf{B}$, resulting in:

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

where $\bar{E} = \bar{\mathbf{c}} \cdot \mathbf{B}$ with $\bar{\mathbf{c}}$ being the mean of the posterior distribution. We provide the explicit form of $\bar{\mathbf{c}}$ in Eq. (25). The expression for $\tilde{\sigma}$ is computationally efficient to evaluate, requiring a single basis evaluation \mathbf{B} followed by $K+1$ dot-products with the mean $\bar{\mathbf{c}}$ and committee parameterisations \mathbf{c}^k .

Having introduced E_{HAL} , it remains to specify an uncertainty measure, or stopping criterion, to terminate the dynamics, identifying new training configurations and extending the training database (as used in AL). To that end, we introduce a relative force uncertainty, f_i , which is attractive from a modelling perspective: For instance, liquid and phonon property predictions require vastly different absolute force accuracy but rather similar relative force accuracy, typically on the order of 3–10%. Given the committee forces $F_i^k = -\nabla_i E^k = \mathbf{c}^k \cdot \nabla_i \mathbf{B}$, 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 force prediction as predicted by the mean parameters $\bar{\mathbf{c}}$. Further, ε is a regularising constant to prevent divergence of the fraction. This parameter is specified by the user and should be chosen to be smaller than the typical force magnitude observed during dynamics. In practice it is found that 0.2–0.4 eV/Å generally yields good behaviour.

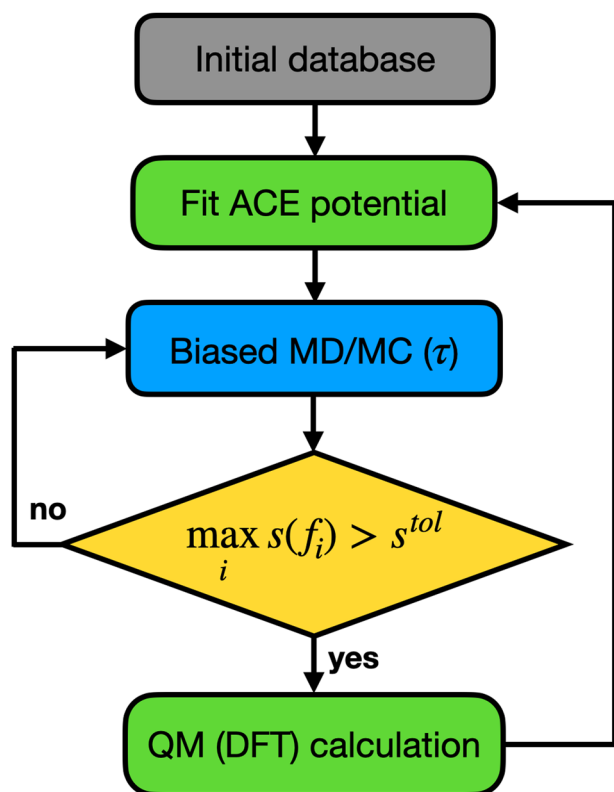


Fig. 1 Schematic overview of the HAL framework performing biased AL controlled by a biasing parameter τ . The softmax normalised relative force uncertainty $s(f_i)$ is used as a stopping criterion to trigger DFT calculations.

During HAL simulations, f_i provides a computationally efficient means to detect emerging local (relative force) uncertainties and is used to 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 requires careful tuning to achieve good performance. Too small f^{tol} keeps triggering unnecessary ab initio calculations, whereas a too large value 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 $s(f_i) = \exp(f_i) / \sum_j \exp(f_j)$, and redefine the stopping criterion as:

$$\max_i s(f_i) > s^{\text{tol}}. \quad (8)$$

This setup is chosen to mimic a probabilistic classifier whereby a binary decision is made between two options: triggering a QM calculation or continue HAL dynamics. A default tolerance of $s^{\text{tol}} = 0.5$ is used as it correspond to the decision boundary between the two options.

The main purpose of this work is to present an accelerated AL scheme for generating data-driven interatomic potentials, and showcase it by determining alloy melting temperature and polymer density with close to experimental accuracy. Using an initial database that comprises a few atomistic configurations, the HAL procedure is started by biasing dynamics towards uncertainty while running MD (and optionally Monte Carlo (MC), for e.g. volume changes or atom swaps). If the uncertainty exceeds a predefined (AL) uncertainty tolerance, or stopping criterion (s^{tol}), during the biased dynamics, a DFT calculation is triggered. This

newly labelled configuration is added to the training database and the ML model is refitted. Then the next HAL iteration commences using the newly fitted potential. The HAL scheme is illustrated in Fig. 1 where τ corresponds to the biasing parameter in Eq. (1). This parameter requires careful tuning and an on-the-fly adaptive scheme controlled by a relative biasing parameter τ_r is discussed in § “Methods”.

The remainder of this article is organised as follows. In the § “Results and discussion” we first demonstrate the suitability of the relative force error measure f_i as a selection criterion in an AL framework by evaluating its correlation with the true relative force error and by using it to sequentially re-assemble a much-reduced diamond structure silicon database (section “Results and discussion”). In sections “AlSi10” and “Polyethylene glycol (PEG)” we show how the HAL framework can be used to build training databases from scratch in the case of an alloy (AlSi10) and polymer (polyethylene glycol or PEG), respectively. The assembled training databases are shown to contain sufficient information to fit ACE potentials that enable stable simulation of MD trajectories on long time scales and can accurately predict macroscopic properties such as the melting temperature of AlSi10 and the density of PEG. Section “Methods” describes the HAL scheme with § “Hyperactive learning (HAL)” describing in detail the Monte Carlo estimate of the HAL biasing force and the adaptive on-the-fly auto-tuning scheme for the determination of the relative biasing parameter τ_r . A brief recap of ACE is also provided in § “Atomic cluster expansion (ACE)” as well as detailed description of the Bayesian regression methods that we use to obtain the uncertainty measures in HAL (§ “(Bayesian) Linear regression” to “Posterior predictive distribution”).

RESULTS AND DISCUSSION

Silicon database filtering

Before illustrating the HAL algorithm itself, we first demonstrate the ability of the relative force uncertainty estimate f_i in Eq. (6) to detect true relative force errors. To that end, we will use this 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 silicon GAP model³⁵ and covers 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 trained are linear ACE models that consist of basis functions up to correlation order $v = 3$, 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, 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 $K = 32$.

Prior to training database reduction the ability of the relative force uncertainty estimate f_i to predict relative force error is investigated. Fig. 2a compares the maximum relative force error in 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 that comprise of the entire silicon database (which in total contains 16,708 atomic site neighbour environments). The regularising constant ϵ in Eq. (6) was set to the mean force magnitude as predicted by the mean parameterisation. Both

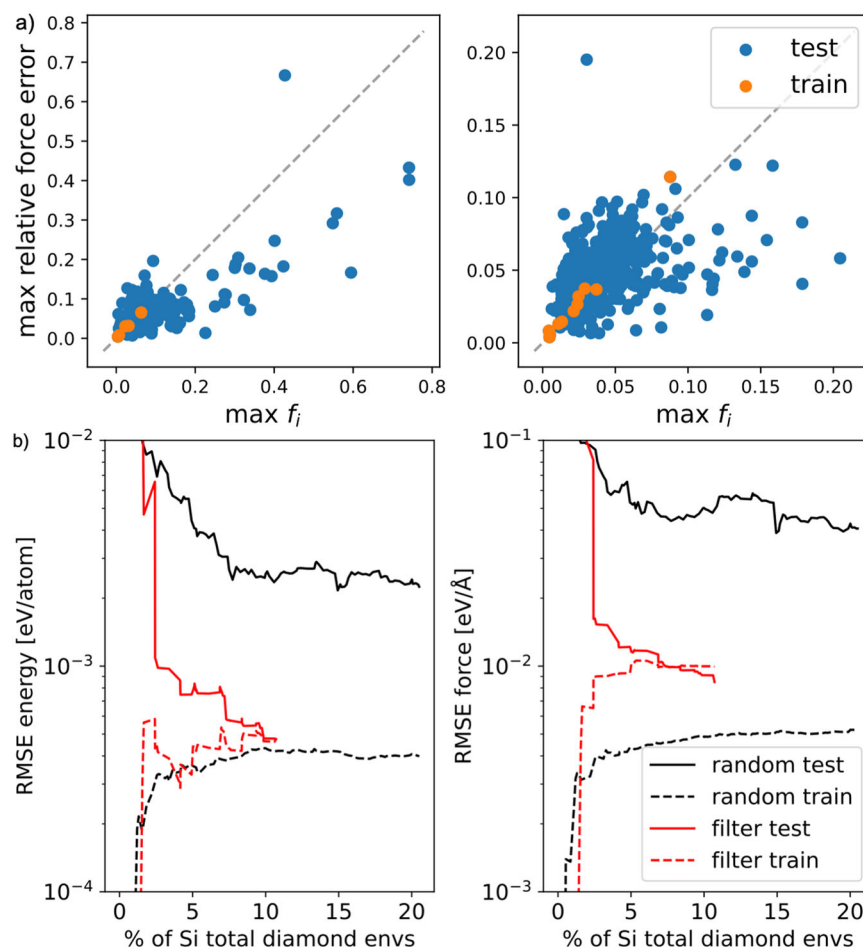


Fig. 2 Benchmarking relative force uncertainty f_i for filtering silicon diamond database. **a** Maximum relative force error estimate $\max f_i$ vs. error correlation plots for silicon diamond containing 4 (left) and 10 (right) training configurations. **b** Learning rate benchmark comparing filtering and random selection for silicon diamond energy (left) and forces (right).

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 algorithms strategy. To demonstrate this, a randomly selected single configuration from the 489 silicon diamond configurations of the silicon database was fitted. Next, f_i was determined over the remaining configurations and the configuration containing the largest $\max f_i$ was added to the training database. This process was repeated. The train and test error for both the energies and forces during this silicon diamond filtering procedure are shown in Fig. 2b. It is benchmarked against performing random selection whereby, starting from the same initial configuration, configurations were chosen at random from the pool of remaining configurations of the training database. 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.

The significant acceleration of the learning rate shown in Fig. 2b shows that generalisation between train and test error is rapidly achieved, in turn suggesting that property convergence is accelerated too. This is investigated by comparing macroscopic properties of the DFT reference with predictions of the

ACE models that were fitted as part of the filtering process. These macroscopic properties include elastic constants, energy volume curves, phonon spectrum and thermal properties for bulk silicon diamond. Results are reported for the ACE models that were fitted to 9 configurations (424 environments), 13 configurations (460 environments) and 17 configurations (608 environments), which, respectively, amount to ~3, 4 and 5% of silicon diamond environments contained in the original database.

Figure 3 demonstrates that property convergence for the energy volume curves, phonon spectrum and thermal properties are rapidly achieved by fitting to a fraction of the original database. Most notably the negative thermal expansion is reproduced, as observed experimentally⁵⁷ and by DFT. This property is particularly challenging to get right and many empirical models fail as shown in the original silicon GAP work³⁵. 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 1. The convergence of the phonon spectrum in Fig. 3 is particularly noteworthy as relative errors on the order of a few percent on small forces ~0.01 eV/Å are typically required to accurately recover the phonon spectrum. The fact that such small relative force errors are achieved while fitting on very few data points is a direct consequence of the design of the filter criterion or uncertainty measure f_i .

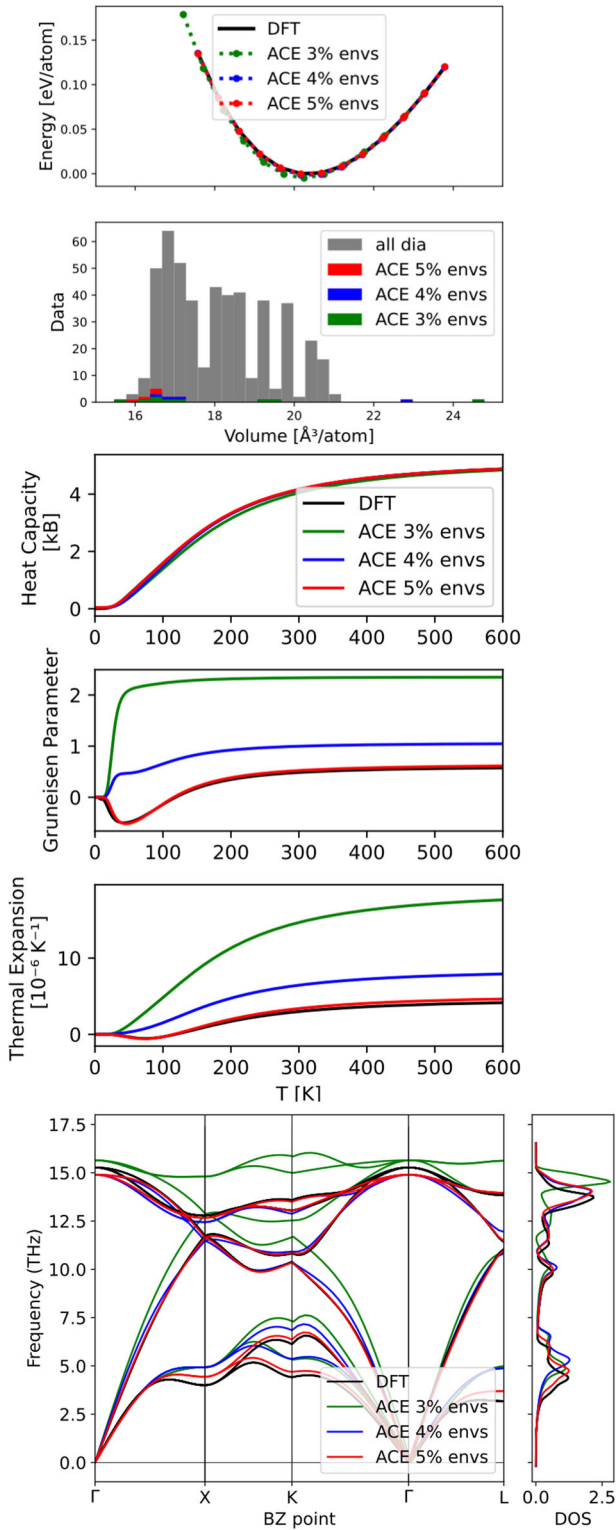


Fig. 3 Property convergence for filtered silicon diamond ACE potentials. Properties included are energy versus volume (top), thermal properties (middle) and phonon spectrum (bottom).

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

Table 1. Convergence of the elastic moduli (GPa) of the filtered ACE models relative to the CASTEP DFT reference.

	B	c_{11}	c_{12}	c_{44}
ACE 3% envs	98.2	188.1	53.3	79.7
ACE 4% envs	84.2	159.8	46.4	75.7
ACE 5% envs	82.5	148.7	49.3	73.7
DFT	82.6	147.2	50.3	73.1

criterion to trigger ab initio evaluations it will be shown how an ACE model is created for AlSi10 using HAL. The ACE models used in this section contained basis functions up to correlation order $\nu = 2$ 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, respectively.

The HAL procedure of building ACE models for alloys starts by creating set of a random crystal structures manually, from which a random alloy and liquid alloy training database are built in an iterative fashion. Once sufficient data for both phases has been collected, the HAL solid and liquid databases are afterwards combined in order to create a model that accurately describes both phases. The first step in the HAL protocol is the creation of a set of small initial random alloy database, which was formed of 32-atom FCC lattice configurations populated with 29 Al and 3 Si atoms, equivalent to 9.7 weight percent Si. This initial random alloy starting database contained ten configurations with lattice constants ranging from 3.80 Å to 4.04 Å and was evaluated using CASTEP⁵⁸ DFT. The main parameters were as follows: plane-wave cutoff 300 eV, kpoint spacing 0.04 Å⁻¹, 0.1 eV electronic 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 $T_{\text{solid}} = 800$ K in order to build the random solid alloy database starting from the 10 initial structures previously described. Besides running biased dynamics, we performed cell volume changes (by adding Gaussian noise to cell vectors) and atom swapping using Monte Carlo (MC) steps during the simulation in order to assist exploration of unseen configurations. These MC steps were accepted or rejected according to the Metropolis-Hastings algorithm⁵⁹.

During HAL dynamics the softmax normalised relative force estimate $s(f_i)$ is evaluated and a ground-truth evaluation triggered once a predefined tolerance of $s^{\text{tol}} = 0.5$ on any of the atoms is met. A total of 42 HAL configurations were sampled as the HAL dynamics at this stage was stable reliably for 5000 steps. The pressure P , temperature T and $\max_i s(f_i)$ are shown in Fig. 4 for four exemplary 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. Finally, as demonstrated in the case of the 43th HAL iteration that increasing the biasing strength to $\tau_r = 0.10$ results in a drastic acceleration (by a factor of 10) in the discovery of configurations with large relative force error.

Next, HAL was employed to assemble a database of liquid random alloys. HAL trajectories were initialised at configurations sampled by cycling through the training database of random solid alloys obtained in the previous HAL run. HAL trajectories were simulated using a Langevin thermostat targeting a temperature regime of $T_{\text{liquid}} = 3000$ K, and a proportional control barostat targeting pressure level of 0.1 GPa. No volume or swap MC steps were performed. After generating 46 liquid alloy configurations

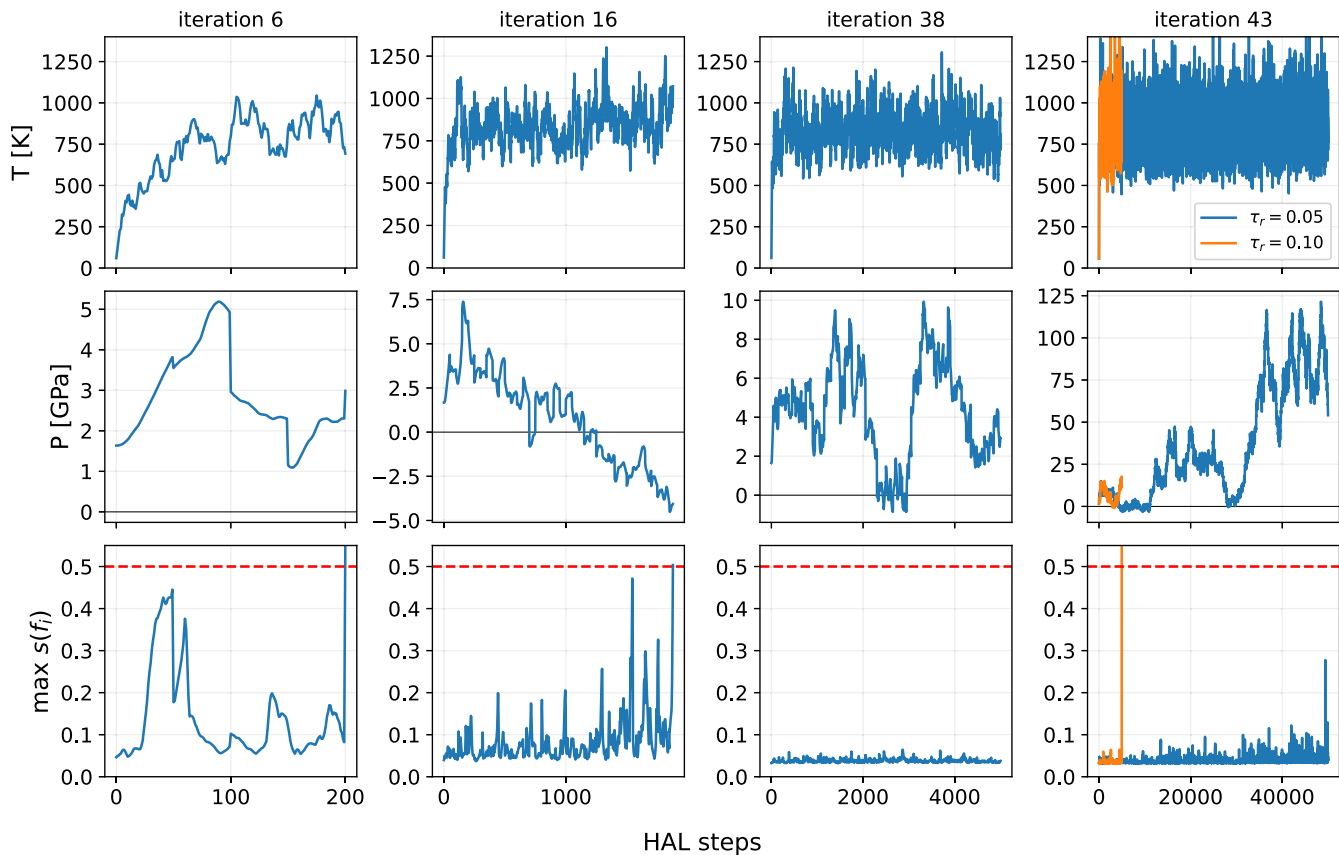


Fig. 4 HAL dynamics for several iterations for the AlSi10 random alloy showing maximum softmax normalised relative force error estimate $\max s(f_i)$, temperature and pressure. DFT calculations are triggered if the tolerance $s^{\text{tol}} = 0.5$ in red is reached. Pressure fluctuations are due to swap/volume MC steps on HAL potential energy surface E_{HAL} .

Table 2. Train/test error splits for HAL generated AISi10 database for varying ARD tolerance α' .															
α'	N_{basis}	Training error		Test error		$N_{\text{atoms}}:$	Evaluation time								Fit time
							$N_{\text{cores}}:$								
		E	F	E	F		1	32	32	448	1	32	32	448	
1k	38	7.693	0.135	8.006	0.147		62	214	72	97	43	403	3	28	2
10k	116	4.199	0.095	6.229	0.104		87	270	100	110	31	319	2	24	3
80k	295	2.401	0.080	5.131	0.089		91	278	105	118	30	310	2	23	17
300k	621	1.869	0.074	5.188	0.095		96	300	115	125	28	287	2	22	63
		(meV/at)	(eV/Å)	(meV/at)	(eV/Å)		(core-μs/atom)				(10 ⁶ step/day)				(s)
Larger ARD tolerance α' includes more basis functions, increases accuracy but leads to worse performance and fitting time. Performance timings for 32 and 14,336 atom (8,8,7 supercell) sized cells are shown for various core counts. A timestep of 1 fs was used such that 10 ⁶ step/day is equivalent to ns/day. These performance timings are for illustrative purposes and do not represent a full computational scaling benchmark, which would need to separately address strong and weak scaling.															

using HAL, the HAL dynamics were reliably stable for 5000 steps and the database assembly for this temperature regime was terminated.

Finally, the 42 HAL generated random alloy configurations and 46 HAL generated liquid configurations were combined to form a training database. This training database was used to fit linear ACE models for AlSi10 using Automatic relevance determination (ARD); see section “Automatic relevance determination (ARD)” for details. We considered various thresholds α' for the pruning of model parameters. The performance of the pruned models in terms of computational speed, training and test errors, are shown in Table

2. The test set used to compute test error consisted of 14 solid and 14 liquid configurations. These configurations were obtained by sampling from the corresponding temperature and pressure regimes by continuing the HAL runs. 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 2 which also includes model evaluation performance and fitting times. Increasing α' not only decreases training error but also test error up to $\alpha' = 300k$ for which the test error increases, a sign of overfitting. Due to the relatively small training database size the computing time to fit the models

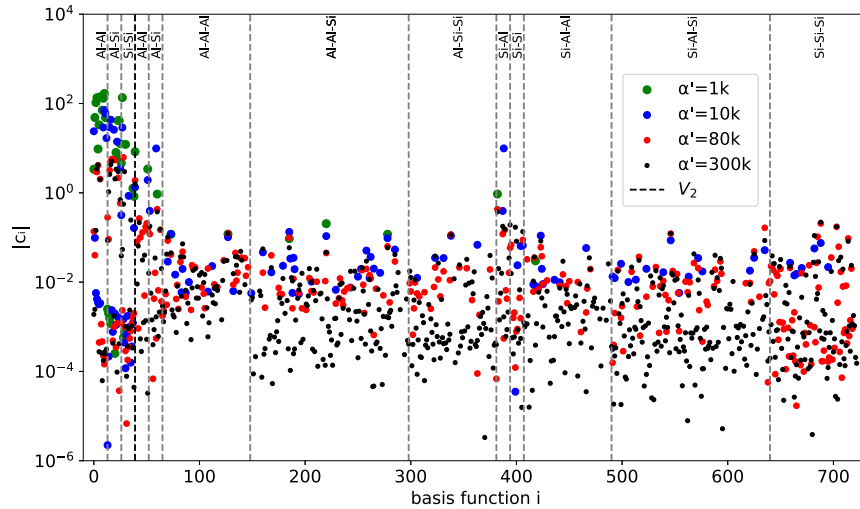


Fig. 5 Coefficient magnitude $|c_i|$ for the 723 basis functions grouped per correlation order and element interaction for various ARD tolerances α' . Large coefficients are assigned to pair interactions, partly captured by the auxiliary pair potential V_2 , as most of the binding energy is contained in these interactions.

remains low, around a minute or less using 8 threads on Intel(R) Xeon(R) Gold 5218 CPU. Performance testing was done using LAMMPS and the PACE evaluator⁶⁰ using Intel(R) Xeon(R) Gold 6142F. The performance tests illustrate scaling trends across different sized cells and cores used for simulation.

Further analysis of the ARD fitted models was done by examining the absolute value of the coefficients $|c_i|$. Basis functions whose estimated prior precision is below the predefined threshold are pruned away as can be seen in Fig. 5. 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 2.

Next, the melting temperature for each of the previously ARD fitted AISi10 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^{61,62}. NS expresses the partition function in term of enthalpy H for N atoms given inverse temperature β , momenta \mathbf{p} and positions \mathbf{q} as follows:

$$\Delta(N, \beta, P) = \int e^{-\beta H(\mathbf{q}, \mathbf{p})} d\mathbf{q} d\mathbf{p} \quad (9)$$

$$\approx \frac{\beta P}{N! h^{3N}} \sum_i w_i e^{-\beta H_i},$$

where the algorithm explores phase space volumes H_i of the PES using a top-down approach, i.e. ideal gas to ground structure. From this expression the heat capacity at constant pressure C_P can be determined:

$$C_P = - \left(\frac{\partial}{\partial T} \frac{\partial \Delta(N, P, \beta)}{\partial \beta} \right), \quad (10)$$

which exhibits a signature peak at a first order phase transition, such as melting. Extensive previous work has shown that NS is a highly automated, efficient, accurate and reliable method for determining the melting temperature without any prior knowledge of the solid phase structure^{63,64}. Because it explores the entirety of configurational space including gas, liquid and solid phases, NS also serves a test for model robustness. This robustness

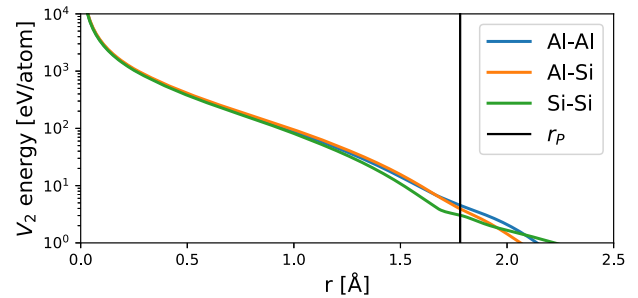


Fig. 6 Pair interaction energy at close approach. Core repulsion is used to stabilise the ACE potential for gas-like configurations explored during NS.

is partly achieved by the addition of the auxiliary repulsive pair potential, V_2 , an example of which is shown in Fig. 6. Core repulsion below spline point r_P is ensured by the addition of a repulsive core shaped $r^{-1}e^{-ar}$, where a is a tuned such that the derivatives across the spline point are smooth⁶⁰.

The NS simulations were carried out using 896 walkers and 32 atom unit cells (29 Al and 3 Si) using the PYMATNEST software⁶⁵. The NS walkers were moved using 1024 steps per NS iteration, each step consisting of MD to move atoms (using a 0.1 fs timestep) and MC for unit cell volume, shearing, stretching and atom-swapping steps, in a ratio of 6:6:6:6, respectively. The pressure was set to 0.1 GPa and the minimum aspect ratio of the unit cell was set to 0.85.

Three independent NS simulations were performed for each of the ACE models fitted to the AISi10 HAL database and the corresponding heat capacity curves shown in Fig. 7. All models predicted the expected fcc ground structure, as confirmed using OVITO's⁶⁶ common neighbour analysis, but a difference in the predicted melting temperature for varying α' can be seen. Only the $\alpha' = 300k$ and $\alpha' = 80k$ models accurately predict the melting temperature of 867 K as given by Thermo-Calc with the TCAL4 database⁶⁷. Comparison with Table 2 suggests that a test accuracy of at least 5 meV/atom is required to determine the melting temperature accurately.

Polyethylene glycol (PEG)

This section presents the application of HAL to build databases for polymers. Polyethylene glycol (PEG) has the formula

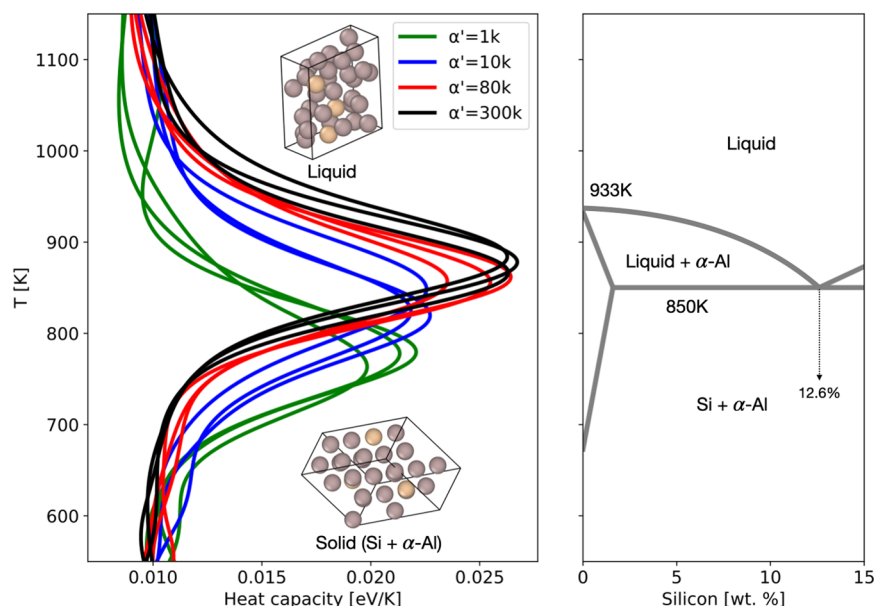


Fig. 7 NS determined heat capacity C_p for ARD fitted linear AlSi10 ACE models (left) and schematic phase diagram for AlSi10⁷⁵ (right). Increasing melting temperature accuracy is demonstrated for fits with large α' .

$\text{H}[\text{OCH}_2\text{CH}_2]_n\text{OH}$, where n is the number of monomer units⁶⁸. 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 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 good accuracy relative to experiment. All DFT reference calculations in this section are carried out with the ORCA code⁶⁹ using the ωB97X DFT exchange correlation functional⁷⁰ and 6-31G(d) basis set.

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($n=2$) polymer configurations was created by running 500 K NVT molecular dynamics simulation using the general purpose ANI-2x forcefield¹¹ sampling structures after every 7000 steps (7 ps) to provide training and test configurations to be used in the following subsections. These ANI-2x sampled configurations were then evaluated using the ORCA DFT using the parameters outlined in the previous paragraph.

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_i were performed using a timestep of 0.5 fs at 500 K. 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 bonds along the chain were within 1.0–2.0 Å and the

CH and OH bonds within 0.8–2.0 Å during the simulation. The minimum number of stable MD timesteps out of the 100 different simulations is shown in Fig. 8 and demonstrates that up to $\tau_i=0.20$ 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 $\tau_i=0.25$ 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 $\tau_i=0.20$, achieves minimum 1 million step MD stability after an order of magnitude fewer exploratory MD timesteps compared to standard AL. This demonstrates that HAL can be used to significantly reduce simulation time required to generate a stable potential, even though a similar amount of training configurations may be required as in a standard AL approach.

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 forcefield¹¹ at 500 K and 800 K using a timestep of 1 fs. Configurations were sampled every 7000 timesteps (7 ps), and used to assemble 500 K and 800 K databases. The 500K database was divided into 750 train configurations and 250 test configurations. The 800 K 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 500 K MD sampled training database HAL was started using $\tau_i=0.10$ and a 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: 500 K, 500 K + 800 K and a 500 K + HAL. Energy scatter plots for these three models are shown in Fig. 9 demonstrating that the errors on the HAL-found configurations are large for both the 500 K and 500 K + 800 K 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

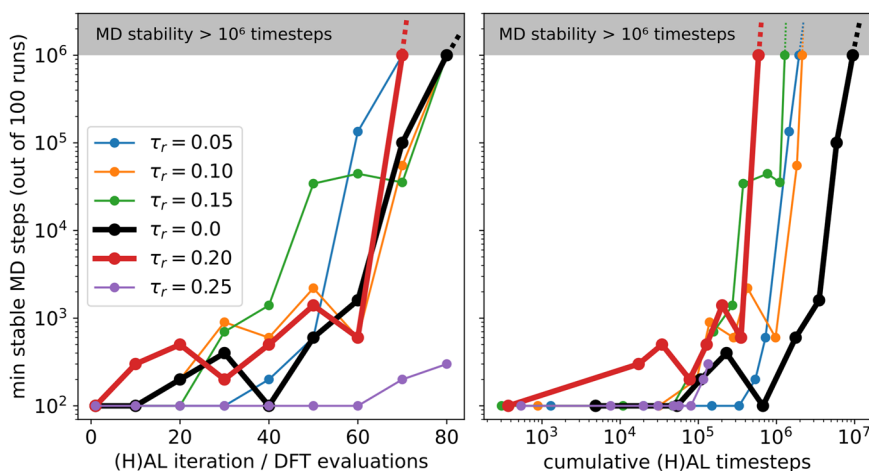


Fig. 8 HAL vs. AL benchmark comparing MD stability for one million MD steps over 100 seeds. Turning on biasing (non-zero τ_r) creates ACE models achieving stable 100 million MD timestep faster than standard AL by up to an order of magnitude.

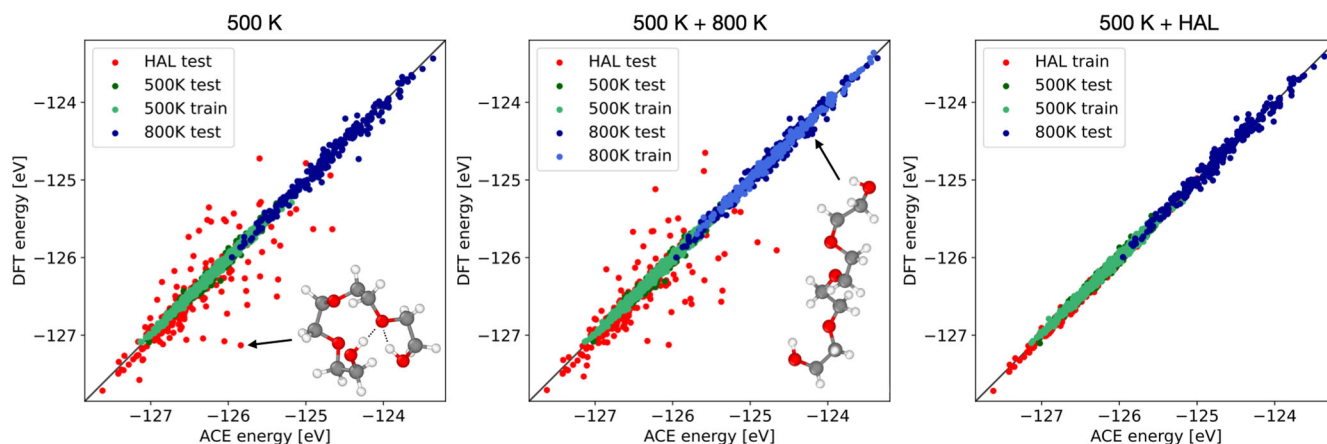


Fig. 9 Energy scatter plots for the 500 K (left), 500 K + 800 K (middle) and 500 K + HAL (right) ACE models. HAL configuration mostly exhibit (double) hydrogen bonding, or rare events, not contained in the MD 500 K/800 K decorrelated samples.

these configurations be reduced as shown in Table 3. Inspection of the HAL generated structures exposes a shared characteristic: most of them contain (double) hydrogen bonding across the polymer an example of which is shown in Fig. 9. 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.

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 in order to prevent fitting on forces too far away from equilibrium.

Table 3. Train and test errors for energies (E) in meV and forces (F) in meV/Å for the 500 K, 500 K + 800 K and 500 K + HAL databases using ACE.

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

^ais test error.

^bis train error.

Using the ACE model a PEG($n=200$) polymer was simulated in LAMMPS⁷¹ with the PACE evaluator pair style with periodic boundary 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

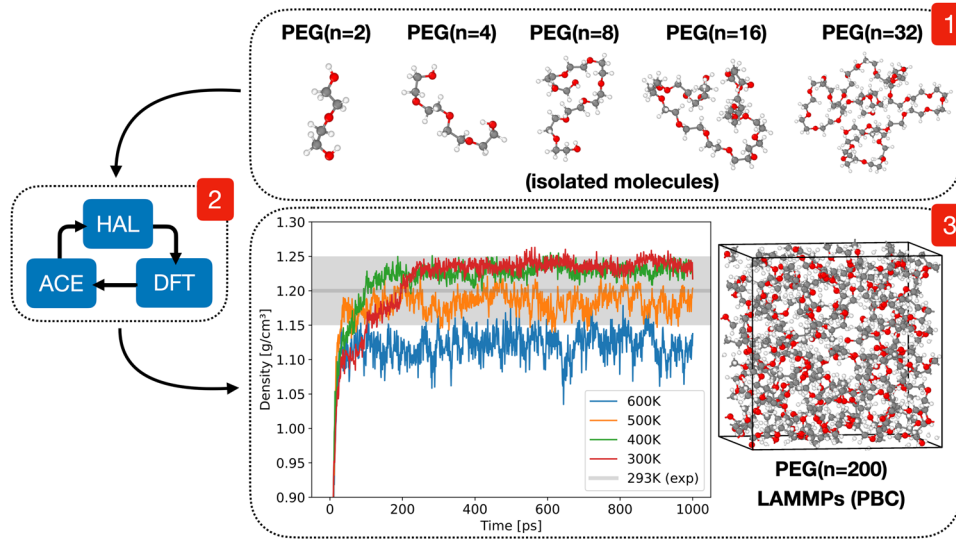


Fig. 10 HAL protocol for building linear ACE PEG model accurately determining PEG($n = 200$) density within experimental accuracy of 1.2 g/cm^3 at 297 K (shaded area)⁷². Training database only included small polymers ranging from $n = 2$ to $n = 32$ in isolation.

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 core- μs /atom per MD step. LAMMPS NPT simulations were performed at 1 bar using a 1 fs timestep at 300 K, 400 K, 500 K and 600 K. The recorded density as a function of simulation time is plotted in Fig. 10. Using the last 500 ps from the 300 K simulation the density was determined to be 1.238 g/cm^3 . This value is around 3% higher than the experimental value of 1.2 g/cm^3 ³⁷².

METHODS

Hyperactive learning (HAL)

The HAL potential energy E_{HAL} as defined in Eq. (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. Considering $\tilde{\sigma}$ defined in Eq. (5), its gradient $\nabla \tilde{\sigma}$ can be computed as:

$$\nabla \tilde{\sigma} = \frac{\nabla \tilde{\sigma}^2}{2\tilde{\sigma}} \quad (11)$$

where

$$\begin{aligned} \nabla \tilde{\sigma}^2 &= \frac{2}{K} \sum_{k=1}^K (E^k - \bar{E}) (\nabla E^k - \nabla \bar{E}) \\ &= \frac{2}{K} \sum_{k=1}^K (E^k - \bar{E}) (\bar{F} - F^k) \end{aligned} \quad (12)$$

and $F^k = -\nabla E^k$, $\bar{F} = -\nabla \bar{E}$. These predictions are obtained by parameterisations $\{\mathbf{c}_k\}_{k=1}^K$, while $\bar{\mathbf{c}}$ is the analytic mean of the posterior distribution as specified in Eq. (25). The K -sum runs over the energy and force predictions from the committee 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:

$$F^{\tilde{\sigma}} := \nabla \tilde{\sigma}. \quad (13)$$

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 generalises AL. Interestingly, previous work

employed a biasing force using a neural network interatomic potential⁷³ 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 history of the latest 100 timesteps δt to estimate the degree of uncertainty (or extrapolation) and adjust the biasing strength accordingly. The biasing strength τ is given by:

$$\tau = \frac{\tau_r \sum_{m=1}^{100} \|\bar{F}(t - m\delta t)\|}{\sum_{m=1}^{100} \|F^{\tilde{\sigma}}(t - m\delta t)\|}, \quad (14)$$

where the relative biasing parameter τ_r is generally set in the range 0.05 to 0.20 (see Fig. 8 for a numerical study). 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.

Atomic cluster expansion (ACE)

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

$$E(\mathbf{c}; R) = \sum_{i \in R} E_i(\mathbf{c}; R). \quad (15)$$

The atomic energies E_i are linear combinations of ACE basis functions, i.e., $E_i(\mathbf{c}; R) = \mathbf{c} \cdot \mathbf{B}_i(R)$. Here, $\mathbf{B}_i(R)$ denotes the evaluation of the ACE basis on the atomic site environment of the i th atom, $\{(\mathbf{r}_{ij}, z_j)\}_{j \in \mathcal{N}_i}$, which consists of relative positions $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and associated chemical elements z_j , denoted by the atomic number, of neighbouring atoms j . In this work it is chosen to project the atomic site environment onto the following single-element basis

function ϕ_{znlm} :

$$\phi_{znlm}(\mathbf{r}_{ij}, z_j) = \delta_{zz_i} R_n(r_{ij}) Y_{lm}(\hat{\mathbf{r}}_{ij}), \quad (16)$$

followed by a pooling operation resulting in features:

$$A_{iznlm} = \sum_j \phi_{znlm}(\mathbf{r}_{ij}, z_j), \quad (17)$$

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

$$\mathbf{A}_{iznlm} = \prod_{t=1}^v A_{iz_t n_t l_t m_t}. \quad (18)$$

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:

$$\mathbf{B}_i = \mathbf{C} \mathbf{A}_i, \quad (19)$$

where \mathbf{C} contains generalised Clebsch-Gordan coefficients; we refer to^{22,23} 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 $v + 1$.

(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:

$$L(\mathbf{c}) = \sum_{R \in \mathbf{R}} w_E^2 |E(\mathbf{c}; R) - \mathcal{E}_R|^2 + w_F^2 |F(\mathbf{c}; R) - \mathcal{F}_R|^2 \quad (20)$$

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

$$\arg \min_{\mathbf{c}} \|\mathbf{y} - \Psi \mathbf{c}\|^2 + \eta \|\mathbf{c}\|^2, \quad (21)$$

where $\Psi \in \mathbb{R}^{N_{\text{obs}} \times N_{\text{basis}}}$ is the design matrix and the observation vector $\mathbf{y} \in \mathbb{R}^{N_{\text{obs}}}$ collects the observations to which the parameters are fitted. Entries in the design matrix and the observations vector corresponding to force observations and observations of virials are scaled by a factor of w_E/w_F and w_V/w_F , respectively, to account for the relative weighting of the penalty terms in (20). 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 models $\epsilon_R^E, \epsilon_R^F$ which give rise to the generative model:

$$\begin{aligned} \mathcal{E}_R &= E(\mathbf{c}; R) + \epsilon_R^E, \\ \mathcal{F}_R &= F(\mathbf{c}; R) + \epsilon_R^F, \end{aligned} \quad (22)$$

for $R \in \mathbf{R}$. This generative model can be written in short-hand form as:

$$\mathbf{y} = \Psi \mathbf{c} + \epsilon, \quad (23)$$

where ϵ is a linear transformation of the error models $\epsilon_R^E, \epsilon_R^F, R \in \mathbf{R}$.

In the context of this work, $\epsilon_R^E, \epsilon_R^F$ model random perturbations of DFT calculations and are assumed to be mainly present due to the locality assumption and DFT convergence properties, e.g. k-point sampling. For simplicity we assume in this work that the entries of the error model ϵ in the generic representation (21) are statistically independent and Gaussian distributed with mean 0 and precision (inverse variance) λ . In terms of the model (22) this assumption implies $\epsilon_R^E \sim \mathcal{N}(0, \lambda^{-1})$, $\epsilon_R^F \sim \mathcal{N}(\mathbf{0}, \mathbf{I} w_E^{-2} w_F^2 \lambda^{-1})$. In principle, extension to other noise models can be made.

The here assumed noise model gives rise to the likelihood function:

$$p(\mathbf{y}|\mathbf{R}, \mathbf{c}, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{N_{\text{obs}}/2} \exp\left\{-\frac{\lambda}{2} \|\mathbf{y} - \Psi \mathbf{c}\|^2\right\} \quad (24)$$

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

$$\begin{aligned} \bar{\mathbf{c}} &= \lambda \Sigma \Psi^T \mathbf{y} \\ \Sigma^{-1} &= \Sigma_0^{-1} + \lambda \Psi^T \Psi. \end{aligned} \quad (25)$$

In the context of this work, having closed form expressions for both these quantities is desirable as it (1) allows for conceptual easy and fast generation of independent samples $\{\mathbf{c}^k\}_{k=1}^K$ from the posterior distribution, and (2) 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 computationally more costly automatic relevance determination (ARD), which we typically use to obtain a final model fit after the data generation is complete.

Bayesian ridge regression (BRR)

In Bayesian ridge regression the covariance of the prior is assumed to be isotropic, i.e.:

$$p(\mathbf{c}|\alpha) = \mathcal{N}(\mathbf{c}|\mathbf{0}, \alpha^{-1} \mathbf{I}), \quad (26)$$

for some hyperparameter $\alpha > 0$, the precision of the prior distribution.

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

$$\ln \pi(\mathbf{c}) = -\frac{\lambda}{2} \|\mathbf{y} - \Psi \mathbf{c}\|^2 - \frac{\alpha}{2} \|\mathbf{c}\|^2 + C, \quad (27)$$

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. (27) with ridge penalty $\eta = \lambda/\alpha$. This shows that the prior naturally gives rise to a regularised solution, keeping coefficient parameters small.

The determination of the hyperparameters α and λ in BRR is achieved by optimising the marginal log likelihood also known as evidence maximisation²⁴. One first defines the evidence function as:

$$p(\mathbf{y}|\alpha, \lambda) = \int p(\mathbf{y}|\mathbf{c}, \lambda) p(\mathbf{c}|\alpha) d\mathbf{c} \quad (28)$$

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:

$$p(\mathbf{y}|\alpha, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{N_{\text{obs}}/2} \left(\frac{\alpha}{2\pi}\right)^{N_{\text{basis}}/2} \int \exp\left\{-\frac{\lambda}{2} \|\mathbf{y} - \Psi \mathbf{c}\|^2 - \frac{\alpha}{2} \|\mathbf{c}\|^2\right\} d\mathbf{c} \quad (29)$$

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:

$$\ln p(\mathbf{y}|\alpha, \lambda) = \frac{N_{\text{basis}}}{2} \ln \alpha + \frac{N_{\text{obs}}}{2} \ln \lambda - \frac{\lambda}{2} \|\mathbf{y} - \Psi \mathbf{c}\|^2 - \frac{\alpha}{2} \|\mathbf{c}\|^2 + \frac{1}{2} \ln \|\Sigma\| - \frac{N_{\text{obs}}}{2} \ln(2\pi) \quad (30)$$

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

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:

$$p(\mathbf{c}|\alpha) = \mathcal{N}(\mathbf{c}|\mathbf{0}, \mathcal{A}^{-1}) \quad (31)$$

$$\mathcal{A} = \text{diag}(\alpha_1, \dots, \alpha_{N_{\text{basis}}}).$$

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.

Posterior predictive distribution

A key property of the Bayesian approach is that it provides a way to quantify uncertainty of model predictions 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 example, the probabilistic description of the predicted energy E^* at a configuration R^* is:

$$E^* = E(\mathbf{c}; R^*) + \epsilon_{R^*}^E, \quad (32)$$

$$\epsilon_{R^*}^E \sim \mathcal{N}(0, \lambda^{-1}),$$

$$\mathbf{c} \sim \pi(\mathbf{c}).$$

Thus, the posterior predictive distribution of energy, i.e., the conditional distribution $p(E^*|R^*)$, can be verified to be normal:

$$p(E^*|R^*) = \int p(E^*|R^*, \mathbf{c}) \pi(\mathbf{c}) d\mathbf{c} \quad (33)$$

$$= \mathcal{N}(E^*|\bar{\mathbf{c}} \cdot \mathbf{B}, \sigma^2),$$

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

Closed forms of the predictive distribution of other quantities that are linear transformations of the coefficients \mathbf{c} and the noise model can be similarly derived. For quantities that are non-linear and potentially only implicitly defined transformations, approximations of their predictive distribution can be obtained by propagation of the Monte Carlo samples $\{\mathbf{c}_k\}$.

DATA AVAILABILITY

The code, potentials and databases used to generate these potentials can be found on the ACEHAL github page <https://github.com/ACEsuit/ACEHAL>.

Received: 7 November 2022; Accepted: 2 August 2023;

Published online: 13 September 2023

REFERENCES

1. Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, 1902765 (2019).
2. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
3. Keith, J. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
4. Thompson, A. P. et al. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
5. Wang, H., Zhang, L., Han, J. & E, W. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
6. Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. The MLIP package: moment tensor potentials with MPI and active learning. *Mach. Learn. Sci. Technol.* **2**, 025002 (2020).
7. Sossa, G. C., Deringer, V. L., Elliott, S. R. & Csányi, G. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Mol. Simul.* **44**, 866–880 (2018).
8. Deringer, V. L. et al. Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
9. Kapil, V. et al. The first-principles phase diagram of monolayer nanoconfined water. *Nature* **609**, 512–516 (2022).
10. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
11. Devereux, C. et al. Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *J. Comput. Chem.* **16**, 4192–4202 (2020).
12. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
13. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quant. Chem.* **115**, 1051–1057 (2015).
14. Bartók, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, 1–9 (2017).
15. Christensen, A. S., Bratholm, L. A., Faber, F. A. & Anatole Von Lilienfeld, O. FCHL revisited: faster and more accurate quantum machine learning. *J. Chem. Phys.* **152**, 044107 (2020).
16. Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K. R. & Tkatchenko, A. sGDML: constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **240**, 38–45 (2019).
17. Braams, B. J. & Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **28**, 577–606 (2009).
18. van der Oord, C., Dussan, G., Csányi, G. & Ortner, C. Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials. *Mach. Learn. Sci. Technol.* **1**, 015004 (2020).
19. Allen, A. E. A., Dussan, G., Ortner, C. & Csányi, G. Atomic permutationally invariant polynomials for fitting molecular force fields. *Mach. Learn. Sci. Technol.* **2**, 025017 (2021).
20. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
21. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
22. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
23. Dussan, G. et al. Atomic cluster expansion: completeness, efficiency and stability. *J. Comput. Phys.* **454**, 110946 (2022).
24. Kovács, D. P. et al. Linear atomic cluster expansion force fields for organic molecules: beyond RMSE. *J. Chem. Theory Comput.* **17**, 7696–7711 (2021).
25. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 6–13 (2017).
26. Anderson, B., Hy, T. S. & Kondor, R. Cormorant: covariant molecular neural networks. *Adv. Neural Inf. Process. Syst.* **32** <https://proceedings.neurips.cc/paper/2019/hash/03573b32b2746e6e8ca98b9123f2249b-Abstract.html> (2019).
27. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Comput. Chem.* **15**, 3678–3693 (2019).
28. Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *International Conference on Machine Learning* 9377–9388 (2021).

29. Gasteiger, J., Becker, F. & Günnemann, S. GemNet: universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst.* **34**, 6790–6802 (2021).
30. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
31. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. Mace: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
32. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
33. Schütt, K. T. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, 992–1002 (2017).
34. Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G. & Michaelides, A. An accurate and transferable machine learning potential for carbon. *J. Chem. Phys.* **153**, 034702 (2020).
35. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
36. Deringer, V. L., Caro, M. A. & Csányi, G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat. Commun.* **11**, 5461 (2020).
37. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
38. Gubaev, K., Podryabinkin, E. V., Hart, G. L. & Shapeev, A. V. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Comput. Mater. Sci.* **156**, 148–156 (2019).
39. Vandermouse, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 1–11 (2020).
40. Vandermouse, J., Xie, Y., Lim, J. S., Owen, C. J. & Kozinsky, B. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of h/pt. *Nat. Commun.* **13**, 5183 (2022).
41. Sivaraman, G. et al. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Comput. Mater.* **6**, 104 (2020).
42. Wilson, J. T., Hutter, F. & Deisenroth, M. P. Maximizing acquisition functions for Bayesian optimization. *Adv. Neural Inf. Process. Syst.* 9906–9917 (2018).
43. Jørgensen, M. S., Larsen, U. F., Jacobsen, K. W. & Hammer, B. Exploration versus exploitation in global atomistic structure optimization. *J. Phys. Chem. A* **122**, 1504–1509 (2018).
44. Bisbo, M. K. & Hammer, B. Global optimization of atomic structure enhanced by machine learning. *Phys. Rev. B* **105**, 245404 (2022).
45. Merte, L. R. et al. Structure of an ultrathin oxide on pt3sn(111) solved by machine learning enhanced global optimization**. *Angew. Chem. Int. Ed.* **61**, e202204244 (2022).
46. Christensen, M.-P. V., Rønne, N. & Hammer, B. Atomistic global optimization x: a Python package for optimization of atomistic structures. *J. Chem. Phys.* **157**, 054701 (2022).
47. Schwalbe-Koda, D., Tan, A. R. & Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun.* **12**, 5104 (2021).
48. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
49. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566 (2002).
50. Bussi, G., Laio, A. & Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **96**, 090601 (2006).
51. Marsili, S., Barducci, A., Chelli, R., Procacci, P. & Schettino, V. Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. *J. Phys. Chem. B* **110**, 14011–14013 (2006).
52. Dickson, B. M., Legoll, F., Lelievre, T., Stoltz, G. & Fleurat-Lessard, P. Free energy calculations: an efficient adaptive biasing potential method. *J. Phys. Chem. B* **114**, 5823–5830 (2010).
53. Darve, E. & Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **115**, 9169–9183 (2001).
54. Hénin, J. & Chipot, C. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.* **121**, 2904–2914 (2004).
55. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
56. Wehrens, R., Putter, H. & Buydens, L. M. The bootstrap: a tutorial. *Chemometr. Intell. Lab. Syst.* **54**, 35–52 (2000).
57. Okada, Y. & Tokumaru, Y. Precise determination of lattice parameter and thermal expansion coefficient of silicon between 300 and 1500 K. *J. Appl. Phys.* **56**, 314–320 (1984).
58. Clark, S. J. et al. First principles methods using CASTEP. *Z. Kristallogr. Cryst. Mater.* **220**, 567–570 (2005).
59. Chib, S. & Greenberg, E. Understanding the metropolis-hastings algorithm. *Amer. Stat.* **49**, 327–335 (1995).
60. Lysogorskiy, Y. L. et al. Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *npj Comput. Mater.* **7**, 1–12 (2021).
61. Ashton, G. et al. Nested sampling for physical scientists. *Nat. Rev. Methods Primers* **2**, 39 (2022).
62. Pártay, L. B., Csányi, G. & Bernstein, N. Nested sampling for materials. *Eur. Phys. J. B* **94**, 159 (2021).
63. Baldock, R. J., Pártay, L. B., Bartók, A. P., Payne, M. C. & Csányi, G. Determining pressure-temperature phase diagrams of materials. *Phys. Rev. B* **93**, 174108 (2016).
64. Pártay, L. B. On the performance of interatomic potential models of iron: comparison of the phase diagrams. *Comput. Materials Sci.* **149**, 153–157 (2018).
65. Bernstein, N. et al. <https://github.com/libatoms/pymatnest> (2016).
66. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO—the open visualization tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).
67. Tang, M., Pistorius, P. C., Narra, S. & Beuth, J. L. Rapid solidification: selective laser melting of alsi10mg. *JOM* **68**, 960–966 (2016).
68. Karimi, Z., Karimi, L. & Shokrollahi, H. Nano-magnetic particles used in biomedicine: core and coating materials. *Mater. Sci. Eng. C* **33**, 2465–2475 (2013).
69. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).
70. Chai, J.-D. & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106 (2008).
71. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **271**, 108171 (2022).
72. *Polyethylene Glycol [MAK Value Documentation, 1998]*, 248–270 (John Wiley and Sons, Ltd, 2012). <https://onlinelibrary.wiley.com/doi/abs/10.1002/3527600418.mb2532268kske0010>.
73. Schran, C., Brezina, K. & Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **153**, 104105 (2020).
74. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38 (1977).
75. Alghamdi, F. & Haghshenas, M. Microstructural and small-scale characterization of additive manufactured AISi10Mg alloy. *SN Appl. Sci.* **1**, 255 (2019).

ACKNOWLEDGEMENTS

G.C. and C.v.d.O. acknowledge the support of UKCP grant number EP/K014560/1. C.v.d.O. would like to acknowledge the support of EPSRC (Project Reference: 1971218) and Dassault Systèmes UK. C.O. acknowledges support of the NSERC Discovery Grant (IDGR019381) and the NFRF Exploration Grant GR022937. The authors would also like to thank Ioan-Bogdan Magdău for discussions on modelling condensed phase polymers.

AUTHOR CONTRIBUTIONS

C.v.d.O. developed efficient uncertainties for ACE models. M.S. conceived the idea of uncertainty-biasing of AL exploration. C.v.d.O. led the implementation of the HAL framework. D.P.K. helped generate data. C.v.d.O. wrote the first version of the manuscript. All authors discussed the theory and results and edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Cas van der Oord.

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023