

Apax: A Flexible and Performant Framework For The Development of Machine-Learned Interatomic Potentials

Moritz R. Schäfer,^{†,¶} Nico Segreto,^{†,¶} Fabian Zills,[‡] Christian Holm,[‡] and
Johannes Kästner^{*,†}

[†]*Institute for Theoretical Chemistry, University of Stuttgart, Pfaffenwaldring 55, 70569
Stuttgart, Germany*

[‡]*Institute for Computational Physics, University of Stuttgart, Allmandring 3, 70569
Stuttgart, Germany*

[¶]*contributed equally*

E-mail: kaestner@theochem.uni-stuttgart.de

Abstract

We introduce Atomistic learned potentials in JAX (**apax**), a flexible and efficient open source software package for training and inference of machine-learned interatomic potentials. Built on the JAX framework, **apax** supports GPU acceleration and implements flexible model abstractions for fast development. With features such as kernel-based data selection, well-calibrated uncertainty estimation, and enhanced sampling, it is tailored to active learning applications and ease of use. The features and design decisions made in **apax** are discussed before demonstrating some of its capabilities. First, a data set for the room-temperature ionic liquid EMIM⁺BF₄⁻ is created using active learning. It is highlighted how continuously learning models between iterations can

reduce training times up to 85 % with only a minor reduction of the models' accuracy. Second, we show good scalability in a data-parallel training setting. We report that a Gaussian Moment Neural Network model, as implemented in `apax`, achieves higher accuracy and up to 10 times faster inference times than a performance-optimized Allegro model. A recently published Li₃PO₄ dataset, reported with comparable accuracy and inference performance metrics, is used as a point of comparison. Moreover, the inference speeds of the available simulation engines are compared. Finally, to highlight the modularity of `apax`, an equivariant message-passing model is trained as a shallow ensemble and used to perform uncertainty-driven dynamics.

1 Introduction

Traditionally, computational methods to describe the interactions of atoms in chemical systems are grouped in quantum mechanical methods, like density functional theory (DFT), and empirical force fields. While the former group of methods can achieve great accuracy, it is many orders of magnitude slower than much less accurate empirical force fields. The emergence of machine-learned interatomic potentials (MLIPs) has led to a paradigm shift in the field of molecular simulation.^{1,2} Within only a few years, the scope of these potentials has progressed from simple systems requiring thousands of training data points for adequate accuracy^{3,4} to modeling complex systems of tens of thousands of atoms using ever smaller amounts of data.^{5–7} Trained to reproduce the energies and forces of quantum mechanical calculations, MLIPs can closely match the accuracy of the reference methods while achieving linear scaling with the number of particles.

Nowadays, plenty of MLIP codes are publicly available. However, most of them serve primarily as a reference implementation of a particular MLIP model architecture^{6,8,9} and do not aim to be simultaneously modular and flexible method development frameworks, as well as performant training and inference engines. Among these frameworks are SchnetPack,⁴ DeepMDkit,^{10,11} FeNNol¹² and chemtrain.¹³ Schnetpack is designed for highly configurable

model and training workflows and comes with a LAMMPS¹⁴ pair style for all its models. DeepMD features the various DeepMD models^{11,15} and supports distributed molecular dynamics (MD), allowing it to scale to very large systems.¹⁶ The FeNNol package emphasizes building hybrid ML/MM models¹⁷ as well as an integration into the TinkerHP¹⁸ ecosystem. Lastly, chemtrain focuses on advanced training setups, such as coarse-graining and differential trajectory reweighting.¹⁹ The presented work introduces `apax`,²⁰ a Python package for training and applying machine-learned potentials based on the JAX²¹ framework. In contrast to the packages discussed above, the primary methodological focus of `apax` is the iterative generation of datasets, often referred to as learning on the fly or active learning. From a software standpoint, it emphasizes performance and flexibility.

Active learning runs are typically initialized with only a handful of training data points. In an iterative manner, preliminary models perform sampling MD simulations that serve as a pool of candidate configurations. Models are retrained after recomputing the most informative snapshots with a quantum chemistry method. The cycle can be repeated until sufficient model accuracy is achieved.^{22–27}

Typically, transfer learning is used to adapt a pre-trained model to a related dataset, leveraging prior knowledge to improve performance and accelerate convergence on new data.^{28–31} The approach can also be applied to active learning: rather than optimizing all parameters from scratch, parameters from the previous iteration are further refined. Such continuous learning strategies can reduce the necessary number of epochs, thereby lowering overall training costs.

Active learning workflows have various requirements for the methods employed at every step. Utilizing models that can be trained quickly is advantageous, as active learning involves numerous retraining cycles. The fast simulation times of the Gaussian Moment Neural Network (GMNN)^{32,33} model, coupled with its good predictive accuracy, make it an ideal model for active learning workflows and large-scale simulations. The GMNN has been successfully applied to a variety of systems from ionic liquids³⁴ to astrochemical surfaces^{35,36} and

other systems.^{37,38} Nonetheless, `apax` offers modular model abstractions that are compatible with any atom-centered representation. During sampling simulations, it is necessary to estimate the model’s uncertainty and stop trajectories from entering unphysical regimes.³⁹ The shallow ensembles recently proposed by Kellner and Ceriotti,⁴⁰ trained on probabilistic loss functions, offer a reliable way to estimate uncertainties. Additionally, biasing the dynamics towards regions of higher uncertainty can enhance the sampling during those simulations, increasing the informativeness of encountered configurations.^{41,42} A new set of data points needs to be selected once a sampling simulation is completed. These should be, at the same time, informative and sufficiently different from each other. Therefore, `apax` implements the batch data selection methods⁴³ developed in the Kästner group.

The study is structured as follows. We begin by introducing MLIPs, particularly GMNN and the other theoretical methods used in the `apax` package. Afterward, we begin the demonstrations by creating a dataset for the 1-ethyl-3-methylimidazolium tetrafluoroborate ($\text{EMIM}^+\text{BF}_4^-$) room-temperature ionic liquid (RTIL) using active learning. Here, we demonstrate how reusing model parameters from the previous iteration can be used to reduce the training time of the current iteration. Next, we analyze the training performance for the GMNN and equivariant message passing models in a data-parallel setting. The inference performance is investigated for the Atomic Simulation Environment (ASE) and the internal MD engine⁴⁴ for different ensemble kinds and sizes. The performance studies use the $\text{EMIM}^+\text{BF}_4^-$ dataset constructed during active learning and subsequently the Li_3PO_4 dataset by Batzner et al.⁸ where inference speeds are available for the Allegro⁵ and SO3krates⁴⁵ models. Finally, we demonstrate the modularity of `apax`’s design by training an equivariant message-passing model as a shallow ensemble and investigate the sampling advantages offered by uncertainty-driven dynamics (UDD).⁴¹

2 Methods

2.1 Machine Learning Interatomic Potentials

Given an atomic configuration S consisting of coordinates \mathbf{R} and atomic numbers Z , potentials used in MD map from S to a potential energy E . For most MLIPs, the locality of atomic interactions is assumed. As a result, additive functional forms are used, with atomic contributions dependent on their respective local environment.^{1,46}

$$E(S, \theta) = \sum_i^{N_{\text{atoms}}} E_i(\mathbf{G}_i, \theta) \quad (1)$$

Here, θ is the set of parameters that are adjusted during training and \mathbf{G}_i a representation of the local atomic environment around atom i . Individual MLIPs differ in the particular representation of local environments and the regression model used to predict atomic energies. One particular MLIP, developed in the Kästner group, is the GMNN model.^{32,33} The model is composed of the Gaussian Moment descriptor and feed-forward neural networks to compute atomic energies. The descriptor maps the pairwise distance vectors between each atom and its local neighbors to a feature vector which is invariant under translations and rotations of the system. A smooth radial neighborhood density is constructed from a set of basis functions. Many possible basis functions have been proposed in the literature, but here we implement equidistant Gaussians⁴ and the non-orthogonal Bessel-like functions of Kocer et al.⁴⁷ These are linearly combined by element-pair coefficients β_{ij} to form a radial basis R . Angular information is captured by cartesian moments constructed with the distance vectors up to some rotation order L . Pairwise contributions are summed over all neighboring atoms j .

$$\Psi_{i,L,s} = \sum_{j \neq i} R_{Z_i, Z_j, s}(r_{ij}, \beta_{ij}) \hat{\mathbf{r}}_{ij}^{\otimes L} \quad (2)$$

The final descriptor is obtained from a set of full tensor contractions, which are implemented up to $L = 3$ and 4-body terms.

$$\begin{aligned} G_{i,s_1,s_2} &= (\Psi_{i,1,s_1})_a (\Psi_{i,1,s_2})_a \\ &\vdots \\ G_{i,s_1,s_2,s_3} &= (\Psi_{i,1,s_1})_a (\Psi_{i,3,s_2})_{a,b,c} (\Psi_{i,2,s_3})_{b,c} \end{aligned} \quad (3)$$

Atomic energies are predicted from neural networks as $E_i = \text{NN}(\mathbf{G}_i)$ and scaled and shifted by per-element parameters σ_{Z_i} and μ_{Z_i} before being summed up according to Equation (1).

$$E_i = \sigma_{Z_i} \cdot \text{NN}(\mathbf{G}_i) + \mu_{Z_i} \quad (4)$$

In addition to the GMNN model, we also implement an equivariant message passing model, EquivMP, similar to NequIP.⁸

2.2 Uncertainty Quantification

During the sampling simulations of early active learning iterations, it is likely that the model is not yet accurate enough to simulate a stable MD trajectory. It is thus necessary to use stopping criteria for the simulation, which terminate the run before the system explores unphysical configurations. The true deviation from the reference method is not known during a sampling simulation. Therefore, the model's uncertainty, as a natural choice for such an

error estimate, is used as a stopping criterion.

Ensembling multiple independently trained models has proven to be a straightforward way to estimate uncertainties.^{3,25,48} Given the predictions of N_{ens} members, the uncertainty of the ensemble can be obtained from the sample standard deviation σ_x of the predictions.⁴⁹ For conciseness, we use $x \in \{E, F_i\}$ as the equations below can be used for both total energies E and atomic forces F_i with an implied sum over components and atoms for the latter.

$$\sigma_x = \sqrt{\frac{1}{N_{\text{ens}}} \sum_m^{N_{\text{ens}}} (x^{(m)} - \bar{x})^2} \quad (5)$$

Compared to other uncertainty quantification methods,^{50,51} model ensembles typically achieve lower validation errors than a single model.⁵² However, model ensembles face two challenges. First, the training and inference time increase linearly with the number of ensemble members. Second, the estimated uncertainties are usually miscalibrated,^{53,54} i.e., there is a low correlation between estimated uncertainty and actual prediction error.

Instead of ensembling full models, the recently proposed shallow ensembles⁴⁰ obtain multiple model predictions by only ensembling the last neural network layer. As a result, energy uncertainties can be computed at negligible additional cost. The forces can be evaluated in two equivalent ways, either as the gradient of the mean energy or the mean of the Jacobian of the ensemble energy predictions.

$$F_i = \frac{\partial \bar{E}}{\partial r_i} = \frac{1}{N_{\text{ens}}} \sum_m^{N_{\text{ens}}} F_i^{(m)} \quad (6)$$

While the latter is more computationally expensive, it allows computing force uncertainties via Equation (5) and is still cheaper than the equivalent full ensemble. Further, these shallow ensembles are trained on probabilistic loss functions, like the negative loglikelihood

(NLL), or continuous ranked probability score (CPRS),⁵⁵ instead of the typical homoscedastic loss functions such as mean squared error (MSE) and Huber.⁵⁶ Using probabilistic losses ensures that both the means and standard deviations of ensemble predictions follow the empirical distribution.

$$\text{NLL} = \frac{1}{2} \left[\frac{(x - x_{\text{ref}})^2}{\sigma_x^2} + \log 2\pi\sigma_x^2 \right] \quad (7)$$

$$\text{CRPS}_{\mathcal{N}}(|x - x_{\text{ref}}|, \sigma) = \sigma \left\{ \frac{|x - x_{\text{ref}}|}{\sigma} \left[2\Phi\left(\frac{|x - x_{\text{ref}}|}{\sigma}\right) - 1 \right] + 2\varphi\left(\frac{|x - x_{\text{ref}}|}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\} \quad (8)$$

Additionally, it is possible to correct a model’s calibration *post hoc* by scaling the predicted uncertainties by a (sample-dependent) factor determined from a validation set. We refer to the discussion by Kellner and Ceriotti⁴⁰ and Rahimi et al.⁵⁷ for more details on post hoc calibration.

2.3 Molecular Dynamics and Biased Sampling

In atomistic simulations, computing thermodynamic observables requires averaging over configurations sampled from the system’s Boltzmann distribution.^{58,59} Such configurations are typically obtained through MD simulations, which involve the integration of the equations of motion of the atoms in the system. In addition to production simulations, preliminary MLIPs can be used to sample candidate configurations more efficiently than with *ab initio* methods. However, long MD-trajectories are necessary to sample slow degrees of freedom and rare events.^{60–63} So-called enhanced sampling methods can drastically reduce the simulation time required to explore a system’s free energy surface, often by adding a bias E_b to the potential energy. To generate candidate configurations, several enhanced sampling methods have emerged in recent years.^{42,64} One such method is UDD,⁴¹ which uses the predicted uncertainty of MLIP ensembles to drive the dynamics of a system towards regions of high uncertainty.

$$E_b(\sigma_E^2) = A \left[\exp\left(-\frac{\sigma_E^2}{N_{\text{ens}} N_{\text{atoms}} B^2}\right) - 1 \right] \quad (9)$$

The bias potential includes two parameters A and B , which denote the maximal strength of the bias potential and its width, respectively.

2.4 Batch Active Learning

Once a sampling trajectory is completed, new data points can be selected to expand the dataset. Here, Equation (5) could be used as well to choose those configurations with the largest predicted uncertainty. However, doing so in a greedy manner is likely to select adjacent samples from the trajectory, limiting the diversity of selected configurations. Batch active learning (BAL) methods allow incorporation of the similarity of new data points to already chosen ones in the selection process, introducing more diversity into the selection. More concretely, given a pool of data $\mathcal{D}_{\text{pool}} = \{S_1, \dots, S_n\}$, the BAL task consists in finding a batch $\mathcal{D}_{\text{batch}} = \{S_1^*, \dots, S_b^*\} \subset \mathcal{D}_{\text{pool}}$ which maximizes an acquisition function a , dependent on the model parameters.⁶⁵

$$\mathcal{D}_{\text{batch}} = \arg \max_{\{S_1, \dots, S_b\} \subset \mathcal{D}_{\text{pool}}} a(\{S_1, \dots, S_b\}, \boldsymbol{\theta}) \quad (10)$$

The acquisition function is typically constructed in such a way as to ensure diversity between selected data points or other criteria of the underlying data distribution.⁴³ In the present work, we use a greedy maximum distance selection with a last-layer gradient feature map ϕ_{ll} .

$$\phi_{\text{II}}(S) = \nabla_{\theta_{\text{II}}} E(S, \theta) \quad (11)$$

$$S = \arg \max_{S \in \mathcal{D}_{\text{pool}} / \mathcal{D}_{\text{batch}}} \min_{S' \in \mathcal{D}_{\text{pool}} \cup \mathcal{D}_{\text{batch}}} \Delta(S, S') \quad (12)$$

Here, $\Delta(S, S') = \|\phi(S) - \phi(S')\|_2$ is the distance between feature vectors. Equation (12) is applied iteratively, meaning that each application yields the next most distant structure in feature space. For further details, we refer to the original publication.⁴³

3 Software Architecture

The `apax` package is based on the JAX framework,²¹ allowing for GPU acceleration and just-in-time (JIT) compilation. In contrast to other machine learning frameworks, such as PyTorch⁶⁶ and TensorFlow,⁶⁷ JAX is based on principles from functional programming, such as pure functions and composable function transformations. Its functional purity and static computation graphs allow the Accelerated Linear Algebra (XLA) compiler to emit highly optimized code, which usually outperforms other frameworks.⁴⁵ The function transformations enable users to quickly build up complex functions from simpler ones, e.g., transforming a function that returns the potential energy of an atomistic system into one that returns the energy and corresponding forces. `apax` embraces the philosophy of JAX and implements a lot of its functionality in terms of such transformations.

In the remainder of the section, we will discuss the structure of `apax` in more detail. A high-level overview of the data flow and feature availability in the package is displayed in Figure 1. All features of `apax` are exposed to the IPSuite³⁹ workflow manager, co-developed by the authors. It separates the model and deep-learning framework-specific code from general-purpose functionality, such as workflow construction, data versioning, and sharing and model evaluation methods.

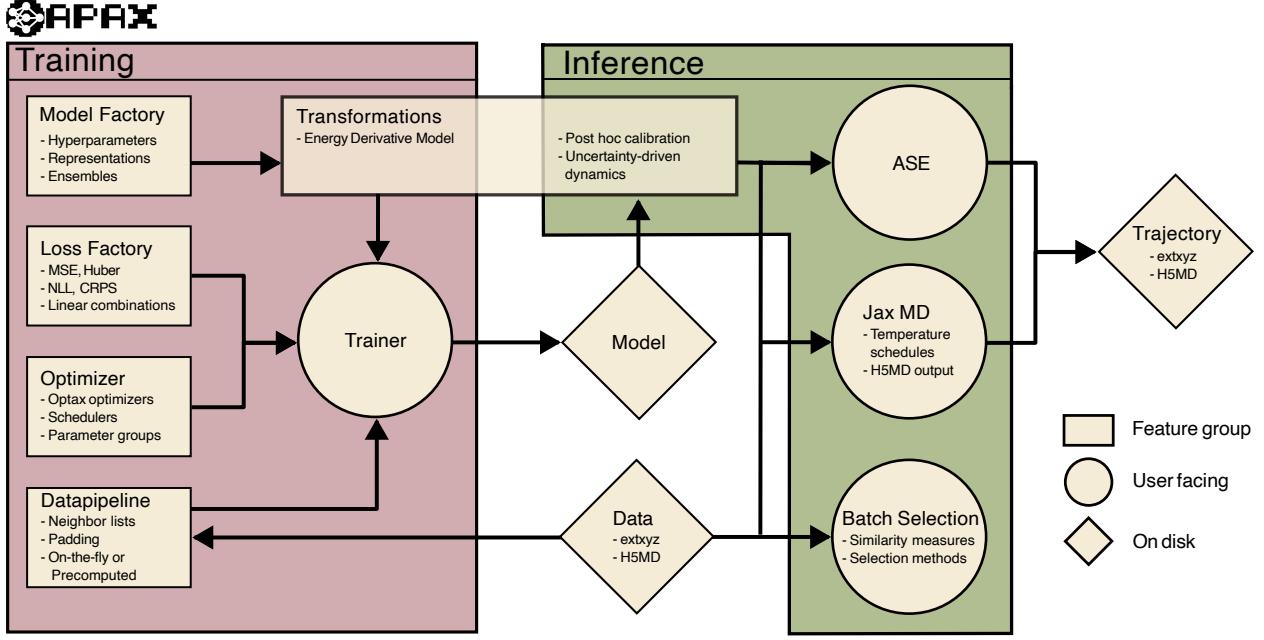


Figure 1: Overview of the features and code structure of the `apax` package. Circles represent user-facing functionalities, rectangles internal feature groups, and diamond shapes data stored on disk.

3.1 Command-Line Interface and Configuration

`apax` provides a command-line interface to utilize its training and JAX-based MD functionalities. To make the initial usage experience as streamlined as possible, the commands `apax template train/md` can be used to create templates for training and MD configuration files. As a file format, we have chosen YAML for its easy human and machine readability. Further, the Pydantic library is used to validate inputs, allowing users to locally check for spelling mistakes of keywords and the correctness of supplied data types before submitting jobs to queuing systems on high performance computers. Some examples of errors that can be caught are given in Listing S3. In the remainder of the section, we illustrate parts of the configuration files where appropriate.

3.2 Data Pipeline

Training MLIPs requires iterating through a collection of training structures hundreds or thousands of times while passing this data to the model for evaluation. The loading of atomistic data requires some careful considerations, especially in a JAX-based framework. First, the training dataset needs to be in an appropriate format. `apax` supports reading atomistic structures into ASE Atoms objects from any ASE readable format, as well as H5MD⁶⁸ files *via* the ZnH5MD⁶⁹ library. The extended XYZ from ASE and H5MD are the most suitable file formats due to their flexibility and, in the latter case, IO performance. As a result, many existing literature datasets can be read directly. Secondly, one needs to account for the static computation graphs required by JAX in the preprocessing of data: All input shapes, e.g., the number of atoms and neighbors, need to be known at compile time, with deviations causing recompilation. While a single recompilation takes only a few seconds, recompiling for every batch in a dataset would be impractically slow. There are two options for handling differently sized systems implemented in `apax`.

For the first option, referred to as “cached”, the systems with the largest number of atoms and neighbors are identified. All samples in the dataset are then padded with zeros to the largest number of atoms, for per-atom quantities such as positions, and the largest number of neighbors, for the neighbor list, respectively. As a result, the training step function only needs to be compiled once, and the additional compute used for the padding is negligible when all structures in the dataset have the same or similar sizes. Instead of padding every sample to the same size, it is also possible to bin system sizes. By allowing for a few recompilations, it is possible to waste less compute on the padding, which is faster for datasets with significantly different system sizes or when a few structures are considerably larger than the median. The bin sizes we have found to work well by default are 10 for the number of atoms and 2000 for the neighbor list. We refer to the second option as “per-batch-padded” or “pbp”. In each case, the preprocessing consists of shuffling the data, computing the neighbor list for all samples in a batch, applying padding, and stacking the input arrays. We find that it is

crucial for efficient MLIP training to compute the neighbor list as part of the preprocessing and not during the training step. `apax` uses the `vesin` library to compute neighborlists.

In the “cached” version, a TensorFlow data pipeline is used to construct a `tf.data.Dataset` from a generator. Here, all samples are processed in the first epoch, and ready samples are cached on disk using `tf.data`. As a result, the data set is not saved in RAM, and batches are loaded asynchronously. In the “pbp” version, we keep a buffer of prepared batches. The buffer is asynchronously filled in a separate thread by a multiprocessing queue, which prepares batches on the fly, allowing the training step and batch preprocessing to overlap.

3.3 Model Abstractions

Separating an ML model implementation into distinct parts increases the modularity and extendability of a framework. The distinct parts are referred to as ‘abstractions’, whereby the central one in `apax` is the `EnergyModel`. Internally, it consists of what we refer to as a representation, readout, and scale-shift layers. To ease the training process, empirical energy correction terms, such as the Ziegler–Littman–Biersack potential,⁷⁰ can be supplied. A representation is any function from pairwise distances, atomic numbers, and the neighbor list to a per-atom feature vector. In `apax`, we currently implement the Gaussian Moment descriptor and an equivariant message passing model similar to NequIP,⁸ based on an example from the e3x⁷¹ documentation. The readout module consists of neural networks and computes per-atom predictions, such as atomic energies, from the previously obtained feature vectors. Notably, shallow ensembling is implemented via the final readout module, and the `EnergyModel` does not sum over the ensemble axis. The final atomic predictions are scaled and shifted by potentially element-dependent, learnable parameters. Common choices for the initialization of these parameters are a least-squares regression³³ or supplying fixed reference values for isolated atom energies.⁷²

Decoupling the representation from distance calculations and energy predictions has two advantages: First, it allows all functionalities in `apax` to be independent of the chosen rep-

resentation, allowing any new representation to instantly access all package capabilities. Second, model developers can focus on the relevant parts that change between architectures. The capabilities of an `EnergyModel` are then expanded by a series of function transformations. For example, the gradient of the energy with respect to positions and the strain tensor allow the prediction of forces and virials, respectively. The calculation of gradients is achieved by utilizing JAX’s automatic differentiation capabilities. Other features such as post hoc calibration, UDD, and the feature maps required by BAL are similarly implemented as function transformations. The model abstractions discussed above are schematically represented in Figure 2.

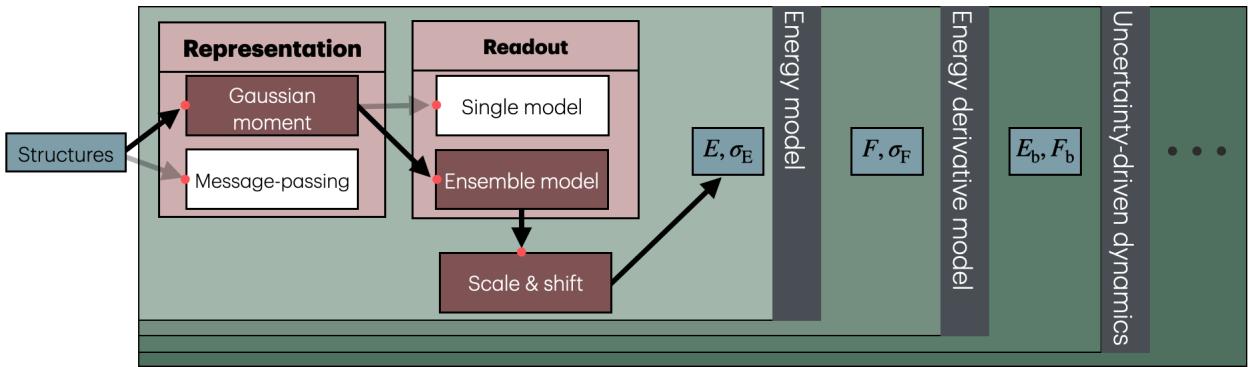


Figure 2: Schematic representation of the model abstractions and interaction with other functionalities in `apax`. Inputs and model outputs added with each transformation are contained in the gray boxes.

3.4 Training

In addition to model- and data pipeline-specific features, the trainer allows for flexible choices of optimizers, learning rates, and loss functions. We interface with Optax,⁷³ a widely used library for stochastic optimization. Optax implements optimizers, such as SGD,⁷⁴ Adam⁷⁵ and SAM,⁷⁶ and we have additionally added the recently proposed AdEMAmix⁷⁷ algorithm. A few Optax optimizers require special treatment in the training step and are not available. In each case, they can be selected from the configuration file, with the option to supply optimizer-specific keyword arguments. Further, `apax` offers a linear decay and a cyclic cosine

learning rate scheduler compatible with any optimizer. Model parameters are grouped to allow separate learning rate selection for neural networks, embedding, and other parameters. The loss factory provides a set of probabilistic and non-probabilistic loss functions, all of which can be used individually or linearly combined. For each loss term, it is possible to set a weighting in the overall loss function and options to divide it by the number of atoms in a structure.

To accelerate training, `apax` makes use of JAX’s sharding API to automatically use all available GPUs or accelerator hardware, known as devices, for data parallelism. For very large models, strategies such as pipeline⁷⁸ and tensor parallelism⁷⁹ are required. These strategies are not implemented as speed requirements in MD simulations, essentially ruling out too large models as impractical. During training, there are several options for tracking metrics. `apax` makes use of callbacks, which allow the logging of training metrics to CSV files, TensorBoard, and MLFlow. Further, when using IPSuite,³⁹ metrics can additionally be monitored using Data Version Control. For transfer learning tasks, `apax` offers an interface to choose which parameters to load (listing S1), which of these to freeze during fine-tuning, and which to re-initialize.

3.5 Deployment

Models trained in `apax` can be applied to various tasks a force field might be used for, such as MD and geometry optimization. For these purposes, we provide an interface to ASE and implement an MD engine, using the neighborlist and thermostats from JaxMD. The ASE calculator provided by `apax` can be used in Python scripts like any other calculator, but was extended with methods for batch evaluation of structures. While the ASE interface is the most flexible, making it particularly useful for quickly setting up custom simulations, the internal MD engine achieves the performance required for production scales. ASE is bottlenecked by blocking host-device transfers happening at every time step and the use of NumPy-based integrators. Such bottlenecks are avoided in the JAX-based MD engine, as

the entire simulation loop takes place on the GPU and is JIT compiled. Further, the transfer of configurations for trajectory writing also happens asynchronously by using JAX’s non-blocking host callbacks, leading to excellent device utilization. Both engines use the neighbor list provided by JaxMD. For periodic systems with very small cell sizes, the minimum image convention used by JaxMD breaks down. Whenever the cell is small enough that multiple images of a neighbor are present in the local environment of an atom, the ASE calculator falls back to using the `vesin` neighbor list.

While the ASE readable formats, such as extended XYZ, are used commonly for distributing datasets and training atomistic machine learning models, they are not optimized for disk size, reading, or writing speed, and are usually fairly restrictive in the kinds of data that can be stored. For example, it is common in MD simulations with MLIPs to calculate the model’s uncertainty during the simulation. Uncertainties computed on the fly can be used to terminate simulations,^{34,80,81} detect rare events,⁸² or compute uncertainties in ensemble averages.⁸³ H5MD, by contrast, is a file format intended for high-performance MD simulations and can be written asynchronously. The ZnH5MD package allows for writing ASE Atoms objects to and reading them from H5MD files. In addition, all properties of the calculator results of the ASE Atoms objects can be stored. They are seamlessly (de)serialized with all other information in both the ASE and internal MD engine.

Equation (6) has a particularly convenient implication for MD simulations: Since it is possible to evaluate the forces of a model cheaply, without associated uncertainties (left-hand side) or more costly, with uncertainties (right-hand side), it is possible to switch between the two modes during a simulation. In the JAX-based MD engine, we implement the switching such that the more expensive force uncertainty evaluation is only performed every N steps, where N is the dump interval of the trajectory. As a consequence, well-calibrated force uncertainties can be calculated at almost no extra cost.

Finally, `apax` comes with two possible ways of distributing models. First, model checkpoints can be uploaded to a cloud storage provider. Second, tight integration with the

ZnTrack⁸⁰ ecosystem allows for the straightforward download of models with a single line of code. Models distributed *via* ZnTrack can be used directly in ZnTrack/IPSuite³⁹ workflows or standalone applications. The models created for the present work are available via `zntrack.from_rev(<model_name>, remote="https://github.com/apax-hub/apax_paper")`.

4 Results and Discussion

4.1 Batch Data Selection

apax augments the batch data selection methods it implements with visualizations to determine how many new data points should be selected. To illustrate the selection process, we train a GMNN model on a 1000-structure subset of the alanine tetra-peptide dataset from the MD22 collection.⁸⁴ We use the resulting model to perform a 10 ps MD simulation of that molecule, followed by a geometry optimization. The data pool is subsequently constructed from the entire optimization trajectory and every 200th configuration from the MD trajectory. 10 samples are selected using maximum distance selection with last-layer features. During the batch selection process, data points are ranked by the squared feature-space distance from every previously selected data point and all training samples. By inspecting the sorted nearest-neighbor distances, it is possible to estimate how many highly informative data points the pool contains. Figure 3a) displays the feature distances for the MD trajectory generated above.

The combined trajectory for the data pool is shown in Figure 3b) with the selected configurations highlighted as red dots. Near convergence, geometry optimizations typically contain many very similar configurations, which the batch data selection method detects automatically. As a result, most configurations are selected from the MD trajectory and only 2 configurations from the geometry optimization, none of them from the very end. Further, the similarity of geometry optimization configurations is visible in the features used during the selection. Figure 3c) displays the first two principal components of the last-layer

features for training and pool configurations, where the geometry optimization configurations are grouped closely together on the right-hand side.

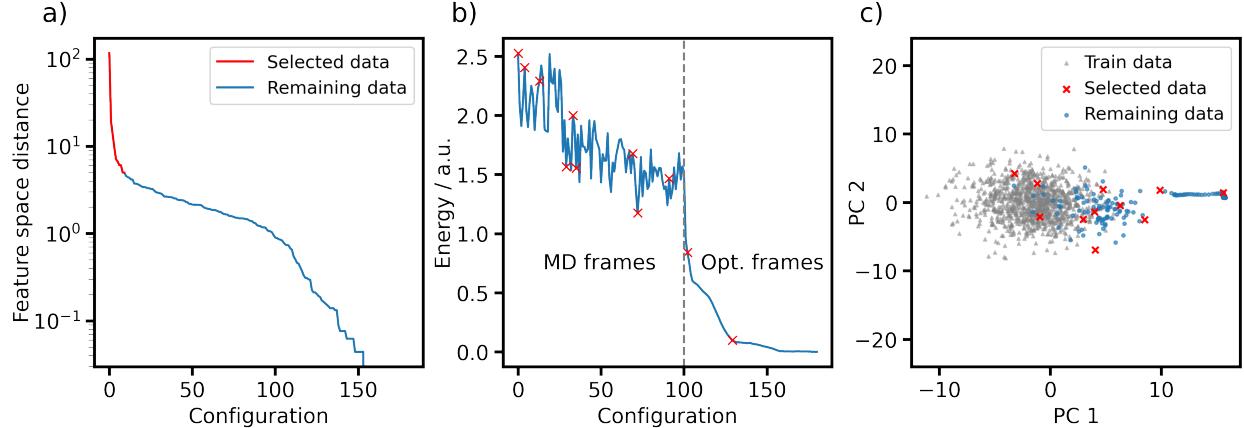


Figure 3: Selection heuristic and analysis of a combined MD and geometry optimization trajectory. a) sorted squared distances used in the MaxDist selection method. b) Energy for each configuration of the combined trajectory. c) First two principal components of the last-layer features for training data and data pool. The pool of data points is marked in blue, selected configurations are marked in red, and training configurations are marked in grey in each subplot.

4.2 Continuous Learning

The following section compares two active learning schemes using batch data selection, focusing on the ionic liquid EMIM⁺BF₄⁻. In the first approach, a new model is initialized and trained from scratch at each iteration. The second is a continuous learning approach in which we retrain the model starting from the parameters of the previous iteration, i.e., a Transfer Learning (TL) step is performed at each iteration. Although more advanced continuous learning techniques are available,^{85,86} this approach represents a robust baseline for such techniques.

Within four active learning workflows, GMNN models are iteratively trained on data generated by the MACE-MP0 foundation model,⁸⁷ which serves as the ground truth for this experiment. The four workflows are initialized with the same core model, trained on the energies and forces of 40 structures, and validated on 20 structures sampled from a MACE-

MP0 trajectory. Every structure consists of 480 atoms with periodic boundary conditions. Subsequently, the models of one workflow are retrained for 1000 epochs at each iteration, referred to as the $R1000_i$. The models of the other workflows underwent TL for 150, 300, and 500 epochs to analyze the effect of the number of epochs in continuous learning approaches. In analogy, they are named $CL150_i$, $CL300_i$, and $CL500_i$. The number i after each workflow label refers to the learning iteration. Besides the number of epochs, all other hyper-parameters are identical across all models. The TL models are trained on all available data. Further, no model parameters are frozen during TL, as early tests revealed that the freezing of parameters degrades performance. More detailed setup is provided in the SI Section S2. Previous studies,³¹ where TL is applied to transfer learned features from a lower quality quantum chemistry method to a higher quality one, found that it is necessary to freeze most of the model weights, retraining only the last or last few layers, to prevent catastrophic forgetting. In contrast to our setup, the dataset size of the target level of theory is often small, and no extended dataset can be used, amplifying such problems.

For each of the four workflows, 240 training and 120 validation structures are accumulated by expanding the training dataset with 20 and the validation dataset with 10 structures over 10 learning iterations. The structures are sampled from 200 ps MD trajectories generated with the models from previous iterations. Sampling simulations are performed with temperature profiles oscillating between 300 K and 600 K to achieve more diversity in the trajectories. As shown in the supporting information Figure S1, the metrics of the models converge with the number of active learning iterations. Comparing $CL150_{10}$, $CL300_{10}$, and $CL500_{10}$ on 150 test structures sampled from a MACE-MP0 trajectory in Figure 4, we found that the model's loss and accuracy metrics decrease with an increasing number of TL epochs. Notably, the force mean absolute error (MAE) of $CL150_{10}$ is already smaller than the force MAE of $R1000_{10}$ at 150 transfer learning epochs. $CL300_{10}$ achieves cumulative savings of 7,000 epochs across 10 iterations while surpassing the accuracy of $R1000_{10}$ on both metrics. Calibrated force uncertainties are monitored throughout simulations. Only $CL300_1$ stopped

after 50 ps due to exceeding an uncertainty threshold of $3.0 \text{ eV} \text{\AA}^{-1}$.

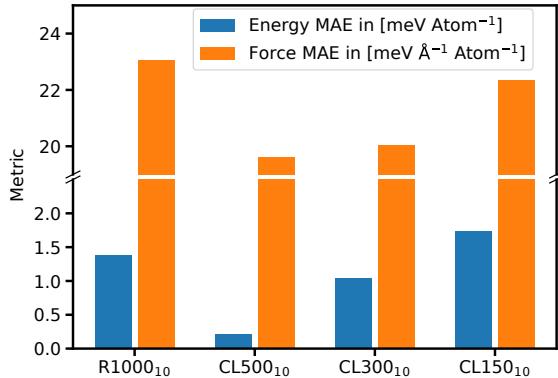


Figure 4: Evaluation metrics of CL150_{10} , CL300_{10} , CL500_{10} and R1000_{10} for a test set of the MACE-MP0 production trajectory.

Further, sampled quantities rather than model metrics are analyzed. The inter-molecular hydrogen-boron radial distribution functions (RDFs) are plotted in Figure 5a) for the final models CL150_{10} , CL300_{10} , CL500_{10} , and R1000_{10} . The RDFs are based on 1 ns simulations of an $\text{EMIM}^+\text{BF}_4^-$ system containing 2400 atoms at 400 K with time steps of 0.5 fs. The difference between the reference RDF and the ones simulated with the models CL150_{10} , CL300_{10} , CL500_{10} and R1000_{10} are displayed in Figure 5b).

Even the least accurate of the four resulting models successfully reproduces the RDFs of the reference potential with a maximal difference of around $\Delta g(r) = 0.05$. The only peaks that are shifted within the accuracy of the bin size of 0.1 Å are the third flat peaks of CL500_{10} and CL150_{10} , with a deviation of -0.1 \AA . For further analysis, we compare the integrated absolute errors (IAE), i.e., the absolute difference between the predicted RDF and the ground-truth RDF integrated over the distance up to 15 Å. The IAEs of the H-B RDFs decreases with an increasing number of transfer learning epochs, also reported in Figure 5b). However, that trend does not hold for some of the atom pairs' RDFs, as shown in Figure S2. The results demonstrate that the continuous learning approach can surpass the accuracy of retrained models for validation metrics and observables while saving considerable training time.

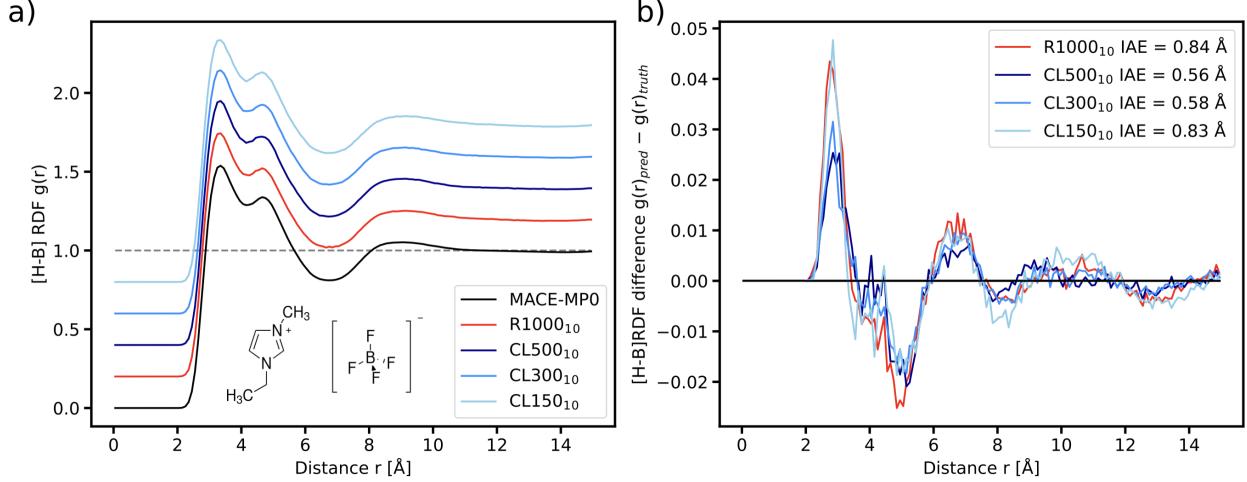


Figure 5: a) Hydrogen-boron RDFs of $\text{EMIM}^+\text{BF}_4^-$ trajectories produced with the final models CL150_{10} , CL300_{10} , CL500_{10} and R1000_{10} and the MACE-MP0 foundation model serving as ground truth. For better visibility, the RDFs are shifted with an offset of 0.2. b) RDFs differences of the final models and MACE-MP0 foundation model.

4.3 Training Parallelization

We investigate the efficiency of the data parallel training strategy by training GMNN and EquivMP models on the $\text{EMIM}^+\text{BF}_4^-$ dataset created in the previous section. Training times are benchmarked with up to 8 GPUs and batch sizes between 8 and 64. The selected range spans from 8 samples to 64, where the former is the minimum batch size capable of utilizing all 8 devices and the latter is the largest power of 2 that fits into the VRAM of a single device for the GMNN model. Both models used a 6.0 Å radial cut-off and 7 radial basis functions. The GMNN model used 5 features for the tensor contractions and two hidden layers with 256 units each. The EquivMP model used a maximal rotation order of $L = 2$, 32 channels, and 2 message-passing steps. All runs are performed on an NVIDIA DGX node with 8 A100 accelerators running CUDA 12.2 and jaxlib 0.4.35. The speed-up in time per epoch compared to training the same model at the same batch size on a single device is displayed in Figure 6. The epoch times are averaged over a training run of 20 epochs, removing the first from the average as it contains the initial compilation time, which is mostly independent of the number of devices. The error bars are the standard error of the

mean calculated from these timings.

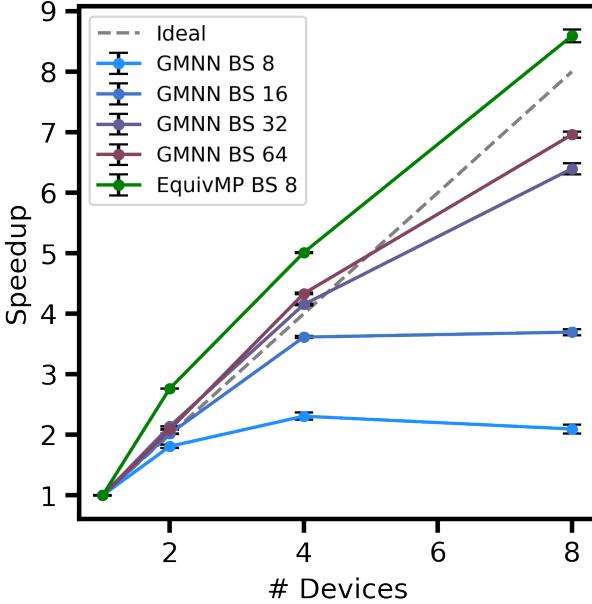


Figure 6: Speed-up of data parallel training compared to a single device for GMNN and EquivMP models for batch sizes between 8 and 64 and 1, 2, 4, and 8 devices.

At a batch size of 8, the GMNN model achieves 90 % of the ideal speed-up on 2 devices. As the number of devices increases at a constant compute load, the communication overhead increases. Since GMNN is an extremely lightweight model, training does not saturate the A100s at low sample sizes per device and the communication overhead is significant, which we also observe for the batch size 16 GMNN model on 8 devices. However, at the larger batch sizes 32 and 64, the scaling remains well across all device counts, with its lowest value being 75 % of the ideal speed-up. In the case of training on 4 devices, we observe a better-than-ideal scaling at these batch sizes. The same behavior can be observed for the EquivMP model for all device counts larger than one. We attribute better-than-ideal scaling behavior to the generation of more efficient GPU kernels by the XLA compiler at smaller sample sizes per device. It should be noted that the expected speedup depends on the system sizes used in the training set and the size of the model. Naturally, for smaller systems and models, the speedup will be smaller than what is reported here.

4.4 Inference Performance

Various ensemble models, along with a single model consisting of one set of parameters, are trained on EMIM⁺BF₄⁻ data to evaluate the inference performance of the JAX-based MD engine and ASE. Additionally, a comparative analysis is conducted between the two available ensemble methods: full and shallow. For each model, five inference times corresponding to different system sizes (480, 960, 1920, 3840, and 7680 atoms) are illustrated in Figure 7. These times are averaged over 50,000 MD steps simulated on an RTX 4090 with a step size of 0.5 fs.

The first kind of model is the full ensemble. For both MD engines, using ensembles with N_{ens} members is more efficient than N_{ens} separate models across all system sizes. Doubling N_{ens} does not result in twice the inference time, instead, it can reduce the step-time per ensemble member by up to 30%, depending on the size of the system and type of ensemble. The best inference time per member reduction compared to a single model is 78% for 480 atoms and $N_{\text{ens}}=8$ members while the least reduction is 7% for $N_{\text{ens}}=2$, resulting in near linear scaling for the largest structure with 7680 atoms.

While the costs of the thermostat and neighbor list updates are constant across ensemble sizes for each system, they are negligible compared to the evaluation of even a single model. Thus, the observed performance gains can be attributed to the generation of more efficient CUDA kernels by the XLA compiler.

The shallow ensemble models with N_{ens} members outperform N_{ens} separate models considerably. The smallest step-time per ensemble member reduction is 60% compared to a single model. As shown in Figure 7a), with the ASE MD engine, they have smaller inference times compared to full ensembles with the same number of ensemble members. With the internal MD engine, see Figure 7b), shallow ensembles can make use of the switching of evaluation modes introduced in Equation (6). Thus, their prediction times are nearly independent of the number of ensemble members N_{ens} , and are, with minor deviations, as fast as a single model. At large system and ensemble sizes, the GPU is saturated, and the

simulation time scales linearly with the number of atoms. For smaller sizes, the scaling is sub-linear. Overall, the internal MD engine consistently outperforms ASE in inference performance. The performance advantage arises primarily due to the JIT compilation of the simulation loop, which unlike ASE is computed entirely on the GPU and the optimizations described in Section 3.5. Performance inconsistencies in ASE inference times are attributed to unoptimized memory transfer processes.

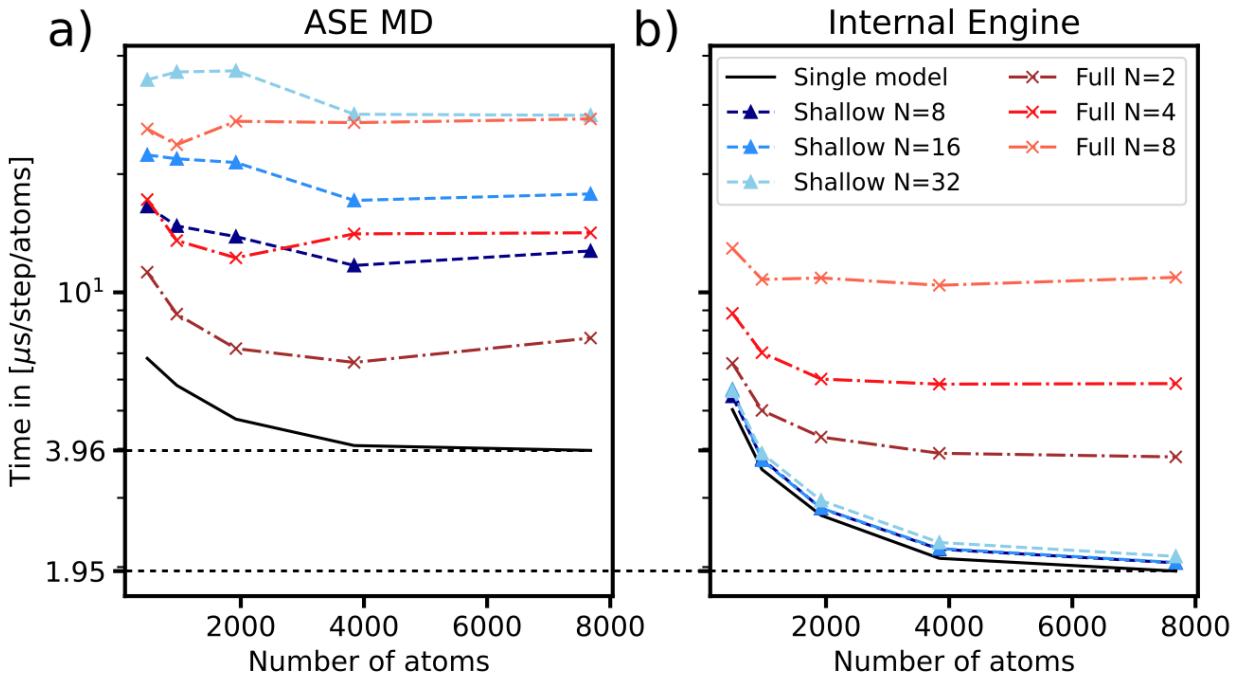


Figure 7: Inference performance of full and shallow ensemble models measure in $\mu\text{s atom}^{-1} \text{step}^{-1}$ for various system sizes. a) timings for the ASE calculator. b) timings for the internal MD engine.

Facilitating comparisons between models and implementations on equal grounds is challenging as performance characteristics vary with the number density of different systems and the hardware used to conduct the simulations.

Hence, we consider the example of the Li₃PO₄ solid-state electrolyte,⁵ for which evaluation metrics and MD simulation speeds of Allegro⁵ and SO3krates⁴⁵ were reported using the same GPU, an Nvidia V100. We follow the training setup of Musaelian et al.⁵ and report energy and force test MAEs as well as the time per MD step in Table 1. The MD simulation

lasted 50 ps and is performed at 600 K. It should be noted that the timings include the entire MD step, not just the model evaluation. Allegro timings were measured in LAMMPS, So3Krates in `mlff`⁸⁸ and GMNN in `apax`' internal MD engine.

Table 1: Test errors and inference times of Allegro, SO3krates and GMNN models trained on a 10K subset of the Li₃PO₄ dataset. The inference time is reported as the wall time per time step per atom for a 192 and a 5154 atom system. All simulations are performed on a Nvidia V100.

	E MAE / meV atom ⁻¹	F MAE / meV Å ⁻¹	#atoms	μs atom ⁻¹ step ⁻¹
Allegro	1.7	73.4	192	27.8
SO3krates	0.2	28.2	192	23.6
GMNN	0.8	68.9	192	10.1
GMNN	0.8	68.9	5184	2.3

In terms of accuracy, GMNN falls between the Allegro and SO3krates models. However, the simulation speed for the 192-atom system is more than 2 times higher than SO3krates and almost 3 times higher than Allegro. It should be noted that in the case of GMNN, the GPU is not saturated for 192 atoms. We repeat the simulation for a 3 × 3 × 3 supercell (5184 atoms) where `apax` reaches 2.3 μs atom⁻¹ step⁻¹, outperforming the other models by more than a factor of 10. The resulting RDF is reported in Figure S3 and shows good agreement with the DFT reference. A common consideration in computational materials science is the accuracy-speed trade-off of various simulation methods. At least for Li₃PO₄ and the observable considered here, all models are in good agreement with the DFT reference. Thus, additional accuracy beyond the Allegro model is not required, while more performant models can significantly reduce the time to solution.

4.5 Uncertainty-Driven Dynamics

Finally, we illustrate the flexibility of `apax` by training an equivariant message passing model as a shallow ensemble and use it to perform UDD. We train an ensemble of 8 shallow members on the alanine tetrapeptide dataset from the MD22 collection.⁸⁴ Two trajectories

are simulated at 300 K using a Langevin thermostat for 50 ps. The first one is an unbiased MD run, while the latter uses a UDD bias following Equation (9) with $A = 1 \text{ eV atom}^{-1}$ and $B = 1.2 \text{ eV}$. To analyze the effect of the enhanced sampling method, we consider the distribution of one dihedral angle during the trajectories. The atoms involved in the dihedral angle and the corresponding histogram are displayed in Figure 8.

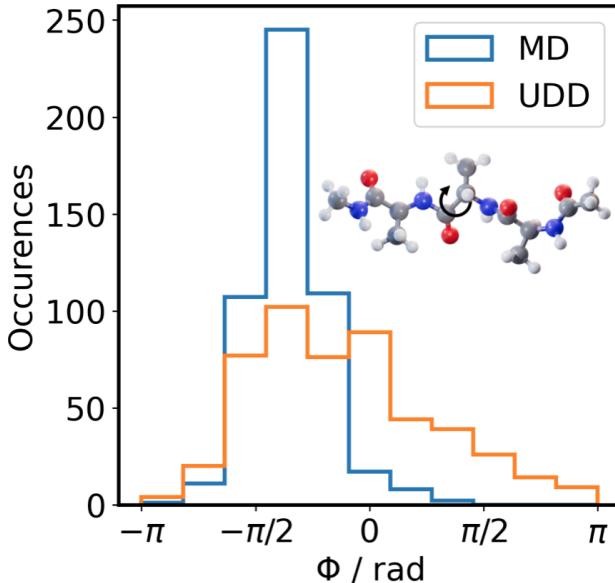


Figure 8: Comparison of alanine tetrapeptide dihedral angle distributions from short unbiased MD and UDD trajectories.

In the unbiased MD run, the dihedral spends most of the time near the energy minimum it started from. The biased trajectory, on the other hand, explores the full rotation around the dihedral angle and can flatten out the histogram significantly, even after such a short simulation.

5 Conclusion

In this work, we presented `apax`, a flexible, easy-to-use, and high-performance framework for training and deploying MLIPs with a focus on active learning. Beyond showcasing its capabilities, it delivers superior performance with streamlined functionality, essential for in silico experiments and rapid method development. It is based on the JAX numerical computing

library, which allows it to run on GPUs and utilize automatic differentiation. The model abstractions implemented in `apax` allow for the straightforward integration of existing JAX-based architectures. Once integrated, they automatically have access to all features in the package, such as batch data selection methods, model ensembling, and UDD. The extensibility of `apax` is achieved through a combination of the powerful function transformations provided by JAX, which are extended by our own.

We have demonstrated the efficacy of `apax` in active-learning scenarios by highlighting several aspects of the package. First, we show how a continuous learning approach can be used to increase model accuracy while simultaneously lowering the number of training epochs compared to training from scratch. Second, in data-parallel settings, `apax` achieves great scaling with the number of devices for GMNN at larger batch sizes and for the equivariant message passing model even at lower ones. For inference, the GMNN model achieves better accuracy than a performance-optimized Allegro while still being between 3 and 12 times faster. Further, when deploying shallow ensembles, the cost of force uncertainty estimation can be made negligible by switching between evaluation modes. By only evaluating force uncertainties on collected configurations, the inference speed of shallow ensembles is essentially identical to a single model. As shallow ensemble uncertainty estimates converge quickly with ensemble size, they only incur a minor training time increase compared to traditional, full ensembles. Finally, we illustrated the flexibility of `apax`'s model abstractions by training an equivariant message-passing model as a shallow ensemble. The model is composed with UDD, which was straightforward to implement as a model-agnostic function transformation. The composed model is used to enhance the configurational sampling of alanine tetrapeptide. In short, `apax` is fast, extandable and well suited for active learning scenarios. It integrates tools like enhanced sampling techniques with selection methods in a straightforward manner. Consequently `apax` is well-suited for studying processes involving slow degrees of freedom, such as slow diffusion of liquids or separated metastable states of biomolecules.

The `apax` package has been successfully used in the study of ionic liquids,³⁴ organic

solvents,³⁹ and CO₂ hydrogenation⁸⁹ with more applications currently being worked on. Nonetheless, there are still opportunities for improvement. While the conditional force uncertainty evaluation of shallow ensembles in the internal MD engine reduces inference time, it does not reduce the memory requirements, as a potentially large ensemble needs to be evaluated in memory. The required memory could be reduced using techniques such as gradient checkpointing or evaluating the model in a loop instead of a vectorized manner. A current limitation is the lack of multi-GPU support for MD. As a result, the maximal system size that can be simulated is limited by the available VRAM on a single GPU. A possible solution is interfacing established MD codes such as Lammmps¹⁴ via chemtrain-deploy⁹⁰ or TinkerHP¹⁸ via Deep-HP.⁹¹ apax is under active development, with these and other additions in progress.

Associated Content

Data and Software Availability

The scripts and workflows needed to reproduce the work presented here can be found at https://github.com/apax-hub/apax_paper. All data generated during the iterative training and production simulations are stored on an S3-object storage. It can be obtained by cloning the git repository and executing `dvc pull` in the repository folder. Further, the data can also be accessed on DaRUS <https://doi.org/10.18419/DARUS-5007>.

All software used throughout this work is publicly available. The `apax` repository is available on Github at <https://github.com/apax-hub/apax> and can be installed from PyPi *via* `pip install apax`. IPSuite is available at <https://github.com/zincware/IPSuite> and can similarly be installed *via* `pip install ipsuite`.

Supporting Information

Configuration file example, validation error output example, training config of active learning experiments, model convergence metrics, EMIM⁺BF₄⁻ RDFs, Li₃PO₄ RDF (PDF)

Author Information

Corresponding Author

Johannes Kästner - Institute for Theoretical Chemistry, University of Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany; <https://orcid.org/0000-0001-6178-7669>; Email: kaestner@theochem.uni-stuttgart.de

Authors

Moritz R. Schäfer - Institute for Theoretical Chemistry, University of Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany; <https://orcid.org/0000-0001-8474-5808>

Nico Segreto - Institute for Theoretical Chemistry, University of Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany; <https://orcid.org/0000-0003-3546-4879>

Fabian Zills - Institute for Computational Physics, University of Stuttgart, Allmandring 3, 70569 Stuttgart, Germany; <https://orcid.org/0000-0002-6936-4692>

Christian Holm - Institute for Computational Physics, University of Stuttgart, Allmandring 3, 70569 Stuttgart, Germany; <https://orcid.org/0000-0003-2739-310X>

Author Contributions

M. R. Schäfer: Conceptualization, Software, Investigation, Visualization, Data Curation, Writing - Original Draft, Writing – Review & Editing

N. Segreto: Conceptualization, Software, Investigation, Visualization, Data Curation, Writing - Original Draft, Writing – Review & Editing

F. Zills: Software, Validation, Data Curation, Writing – Review & Editing

C. Holm: Resources, Writing – Review & Editing, Supervision, Funding acquisition

J. Kästner: Resources, Writing – Review & Editing, Supervision, Funding acquisition

M. R. Schäfer and N. Segreto contributed equally and share first authorship.

Notes

The authors declare no competing financial interests.

Acknowledgements

The authors would like to thank Lisa Schröder for insightful feedback on an early version of the manuscript.

C.H., J.K., F.Z., and M.S. acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the framework of the priority program SPP 2363, “Utilization and Development of Machine Learning for Molecular Applications - Molecular Machine Learning” Project No. 497249646.

N.S. acknowledges support by the Deutsche Forschungsgemeinschaft project number 516238647 - SFB1667/1 (ATLAS - Advancing Technologies for Low-Altitude Satellites), and by the Ministry of Science, Research and the Arts Baden-Württemberg in the Artificial Intelligence Software Academy (AISA).

Further funding through the DFG under Germany’s Excellence Strategy - EXC 2075 - 390740016 and the Stuttgart Center for Simulation Science (SimTech) was provided.

All authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

S1 Configuration file and validation errors

Listing S1 and listing S2 below give examples of the configuration files used for model training and inference using the internal MD engine in `apax`, respectively.

```
checkpoints:  
  # load existing model  
  base_model_checkpoint: path/pre_trained_model  
  # reinitialize scaling and shifting parameters  
  reset_layers: [scale_shift]
```

```
optimizer:  
    emb_lr: 0.0000 # freeze embedding layer  
    # train other parameters  
    nn_lr: 0.0001  
    scale_lr: 0.001  
    shift_lr: 0.0001
```

Listing S1: Sections of an input file relevant for transfer learning demonstrating the loading, freezing and re-initializing of parameters from a previously trained model.

```

ensemble:
  name: nvt
  dt: 0.5 # fs time step
  temperature_schedule:
    name: constant
    T0: <T> # K
  thermostat_chain:
    chain_length: 3
    chain_steps: 2
    sy_steps: 3
    tau: 100

duration: <DURATION> # fs
n_inner: 500 # compiled innner steps
sampling_rate: 10 # dump interval
buffer_size: 100
dr_threshold: 0.5 # Neighborlist skin
extra_capacity: 0

sim_dir: md
initial_structure: <INITIAL_STRUCTURE>
load_momenta: false
traj_name: md.h5
restart: true
checkpoint_interval: 50_000
disable_pbar: false

```

Listing S2: Input file for performing a JAX-based MD simulation with an apax model

apax's command line interface has a built-in validator for its configuration files. The validator is based on Pydantic and can catch missing or misspelled keywords and the usage of incorrect data types. Listing S3 shows the error messages when catching common mistakes in a training configuration file.

```
apax validate train train.yaml
>>> 3 validation errors for config
>>> n_epochs
>>>     Field required
>>>     input_type: int
>>> nepochs
>>>     Extra inputs are not permitted
>>>     input_type: int
>>>     input: 1000
>>> data.directory
>>>     Field required
>>>     input_type: str
>>> Configuration Invalid!
```

Listing S3: Validation errors for a training configuration file in which the number of epochs keyword was misspelled and the training directory was not specified.

S2 Continuous Learning

In listing S4, non-default hyper-parameters used in the active learning experiments are listed.

A full model configuration with all default parameters can be found under

https://apax.readthedocs.io/en/latest/configs/full_configs.html, state Nov. 28, 2024.

```
n_epochs: 1000 / 500 / 300 / 150

data:
    batch_size: 1
    valid_batch_size: 20

model:
    ensemble:
        kind: shallow
        n_members: 16

loss:
- name: energy
    loss_type: crps
- name: forces
    loss_type: crps

optimizer:
    name: adam
    emb_lr: 0.0005
    nn_lr: 0.0005
    scale_lr: 0.0005
    shift_lr: 0.0005
    schedule:
        decay_factor: 0.95
```

Listing S4: Non default hyper-parameter of the models used for the active learning experiments.

Here, we present the MAE and root mean squared error (RMSE) for both energy and force on a validation dataset, plotted against the learning cycles for all four active learning approaches.

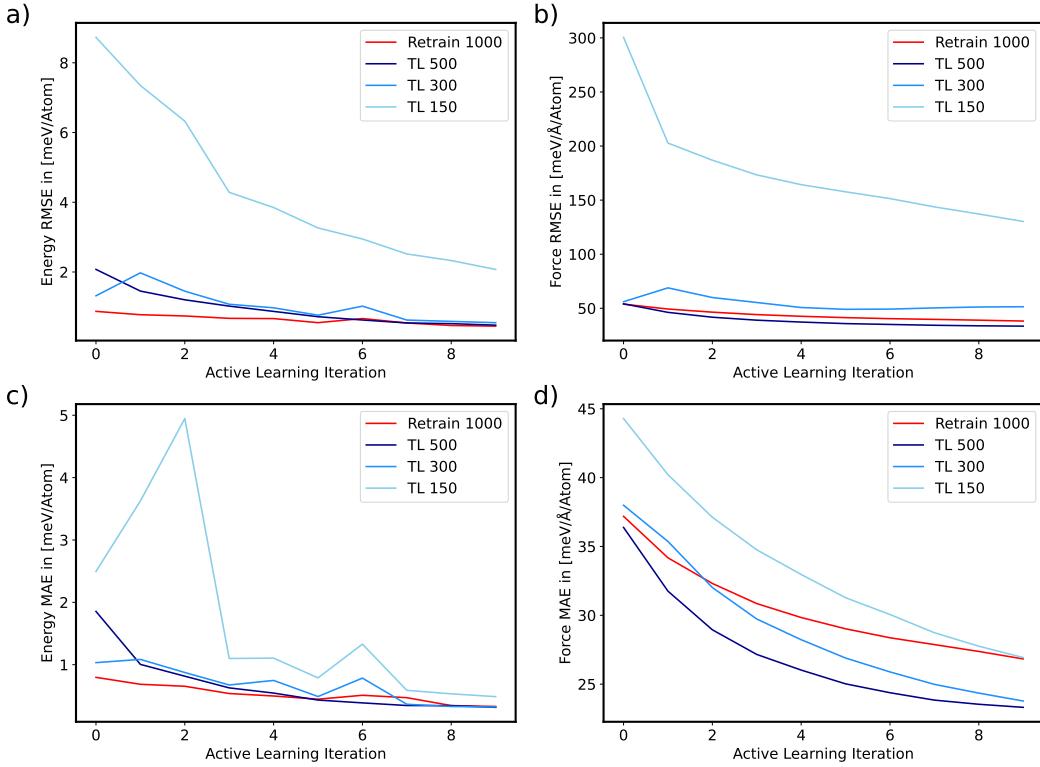


Figure S1: Validation metrics of all 4 active learning approaches plotted against their learning cycles. a) Energy RMSE, b) force RMSE, c) energy MAE, and d) force MAE.

Furthermore, for completeness, all radial distribution functions (RDFs) of $\text{EMIM}^+\text{BF}_4^-$ and their corresponding errors are provided in Figure S2. These are based on foundational trajectories generated using the final models from the four active learning cycles, with the MACE-MP0 model serving as the reference.

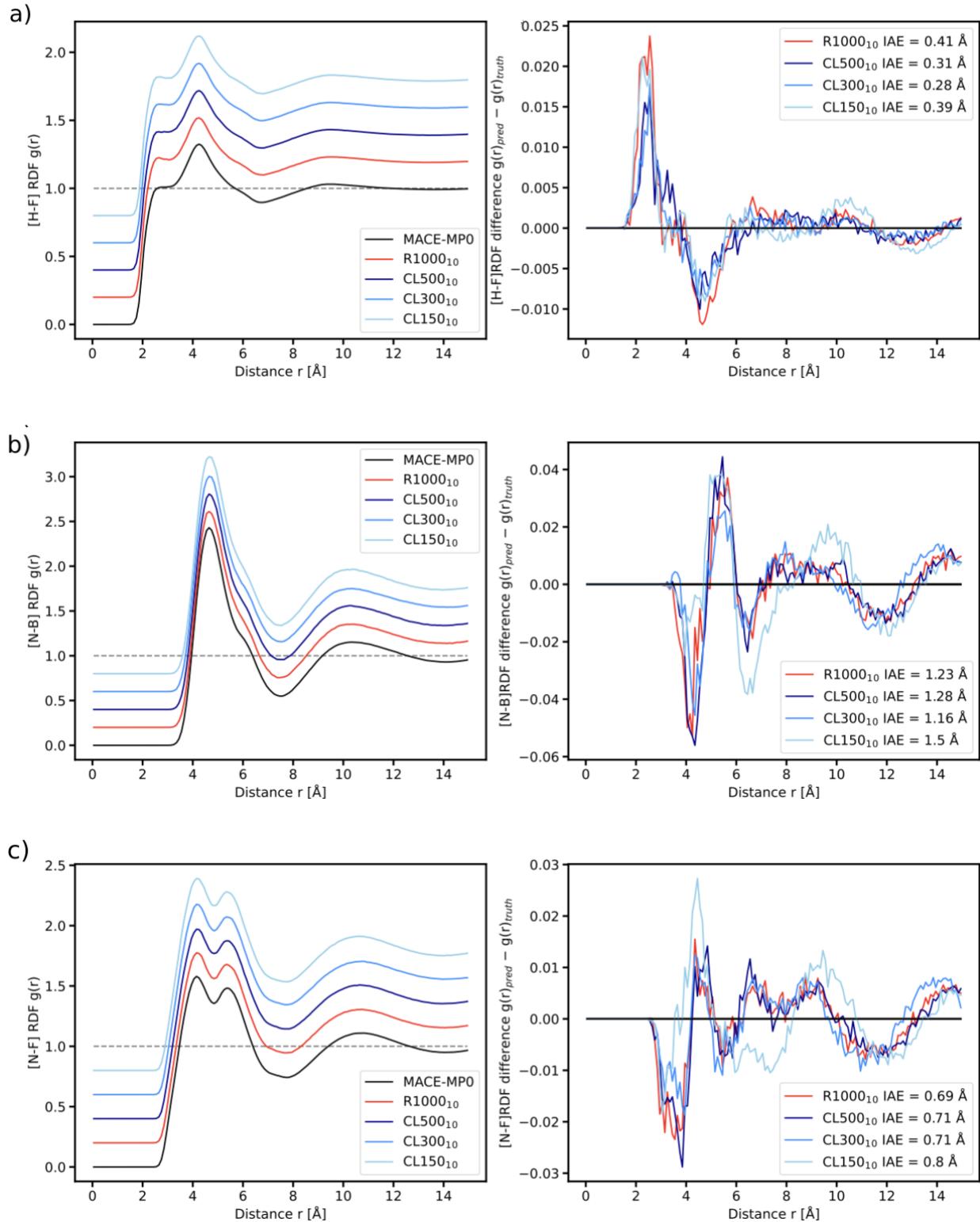


Figure S2: Radial distribution function of $\text{EMIM}^+\text{BF}_4^-$ at 400 K obtained from 1 ns trajectories simulated with MLIPs compared to a MACE-MP0 reference.

S3 Li₃PO₄ Radial Distribution Function

Figure S3 displays the radial distribution functions obtained from the MD simulations of Li₃PO₄ using a GMNN model and a DFT reference. The DFT reference was computed from the 600 K subset of the Li₃PO₄ dataset.⁵ The GMNN model agrees with the reference observable.

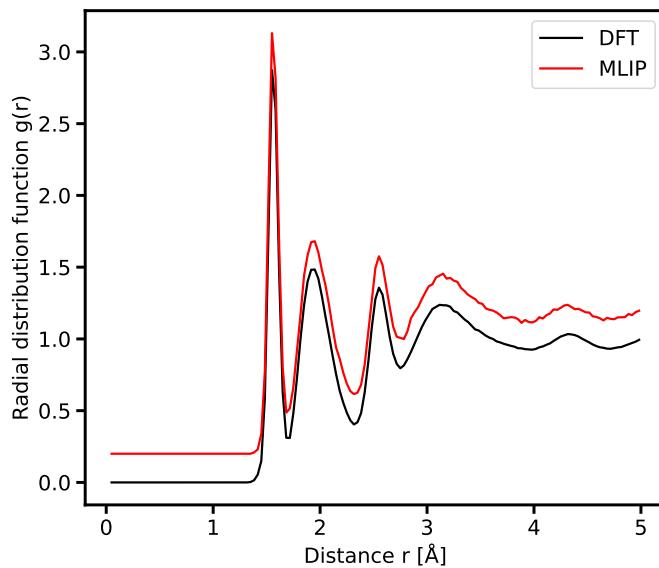


Figure S3: Radial distribution function of Li₃PO₄ at 600 K obtained from a 50 ps trajectory simulated with a MLIP compared to a DFT reference. For visualisation, the MLIP RDF is shifted about 0.2.

References

- (1) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chemical Reviews* **2021**, *121*, 10142–10186.
- (2) Behler, J.; Csányi, G. Machine Learning Potentials for Extended Systems: A Perspective. *The European Physical Journal B* **2021**, *94*, 142.

- (3) Artrith, N.; Behler, J. High-Dimensional Neural Network Potentials for Metal Surfaces: A Prototype Study for Copper. *Physical Review B* **2012**, *85*, 045439.
- (4) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.
- (5) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics. *Nature Communications* **2023**, *14*, 579.
- (6) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *Advances in Neural Information Processing Systems* **2022**, *35*, 11423–11436.
- (7) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-Fly Active Learning of Interpretable Bayesian Force Fields for Atomistic Rare Events. *npj Computational Materials* **2020**, *6*, 1–11.
- (8) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nature Communications* **2022**, *13*, 2453.
- (9) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* **2019**, *15*, 3678–3693.
- (10) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Computer Physics Communications* **2018**, *228*, 178–184.

- (11) Zeng, J. et al. DeePMD-kit v2: A Software Package for Deep Potential Models. *The Journal of Chemical Physics* **2023**, *159*, 054801.
- (12) Plé, T.; Adjoua, O.; Lagardère, L.; Piquemal, J.-P. FeNNol: An efficient and flexible library for building force-field-enhanced neural network potentials. *The Journal of Chemical Physics* **2024**, *161*, 042502.
- (13) Fuchs, P.; Thaler, S.; Röcken, S.; Zavadlav, J. Chemtrain: Learning Deep Potential Models via Automatic Differentiation and Statistical Physics. 2024; <http://arxiv.org/abs/2408.15852>.
- (14) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Computer Physics Communications* **2022**, *271*, 108171.
- (15) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. End-to-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended Systems. Advances in Neural Information Processing Systems. 2018.
- (16) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Physical Review Letters* **2018**, *120*, 143001.
- (17) Plé, T.; Lagardère, L.; Piquemal, J.-P. Force-Field-Enhanced Neural Network Interactions: From Local Equivariant Embedding to Atom-in-Molecule Properties and Long-Range Effects. *Chemical Science* **2023**, *14*, 12554–12569.
- (18) Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.;

- Ponder, J. W.; Piquemal, J.-P. Tinker-HP: A Massively Parallel Molecular Dynamics Package for Multiscale Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields. *Chemical Science* **2018**, *9*, 956–972.
- (19) Montes-Campos, H.; Carrete, J.; Bichelmaier, S.; Varela, L. M.; Madsen, G. K. H. A Differentiable Neural-Network Force Field for Ionic Liquids. *Journal of Chemical Information and Modeling* **2022**, *62*, 88–101.
- (20) Schäfer, M. R.; Segreto, N.; Zills, F. Apax-Hub/Apax: V0.7.0. 2024; <https://zenodo.org/records/14002846>.
- (21) Bradbury, J.; Frostig, R.; Hawkins, P.; Johnson, M. J.; Leary, C.; Maclaurin, D.; Necula, G.; Paszke, A.; VanderPlas, J.; Wanderman-Milne, S.; Zhang, Q. JAX: composable transformations of Python+NumPy programs. 2018; <http://github.com/jax-ml/jax>.
- (22) Cohn, D. A. Neural Network Exploration Using Optimal Experiment Design. *Neural Networks* **1996**, *9*, 1071–1083.
- (23) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Physical Review Letters* **2015**, *114*, 096405.
- (24) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science* **2017**, *140*, 171–180.
- (25) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics* **2018**, *148*, 241733.
- (26) Shuaibi, M.; Sivakumar, S.; Chen, R. Q.; Ulissi, Z. W. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Machine Learning: Science and Technology* **2020**, *2*, 025007.

- (27) Zaverkin, V.; Kästner, J. Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design. *Machine Learning: Science and Technology* **2021**, *2*, 035009.
- (28) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345–1359.
- (29) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* **2019**, *10*, 2903.
- (30) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data* **2020**, *7*, 134.
- (31) Zaverkin, V.; Holzmüller, D.; Bonfirraro, L.; Kästner, J. Transfer Learning for Chemically Accurate Interatomic Neural Network Potentials. *Physical Chemistry Chemical Physics* **2023**,
- (32) Zaverkin, V.; Kästner, J. Gaussian Moments as Physically Inspired Molecular Descriptors for Accurate and Scalable Machine Learning Potentials. *Journal of Chemical Theory and Computation* **2020**, *16*, 5410–5421.
- (33) Zaverkin, V.; Holzmüller, D.; Steinwart, I.; Kästner, J. Fast and Sample-Efficient Interatomic Neural Network Potentials for Molecules and Materials Based on Gaussian Moments. *Journal of Chemical Theory and Computation* **2022**, *17*, 6658–6670.
- (34) Zills, F.; Schäfer, M. R.; Tovey, S.; Kästner, J.; Holm, C. Machine learning-driven investigation of the structure and dynamics of the BMIM-BF4 room temperature ionic liquid. *Faraday Discussions* **2024**, *253*, 129–145.

- (35) Molpeceres, G.; Zaverkin, V.; Kästner, J. Neural-network assisted study of nitrogen atom dynamics on amorphous solid water – I. adsorption and desorption. **2020**, *499*, 1373–1384.
- (36) Molpeceres, G.; Zaverkin, V.; Watanabe, N.; Kästner, J., Binding energies and sticking coefficients of H₂ on crystalline and amorphous CO ice. *A&A* **2021**, *648*, A84.
- (37) Gubaev, K.; Zaverkin, V.; Srinivasan, P.; Duff, A. I.; Kästner, J.; Grabowski, B. Performance of Two Complementary Machine-Learned Potentials in Modelling Chemically Complex Systems. *npj Computational Materials* **2023**, *9*, 1–15.
- (38) Zaverkin, V.; Netz, J.; Zills, F.; Köhn, A.; Kästner, J. Thermally Averaged Magnetic Anisotropy Tensors via Machine Learning Based on Gaussian Moments. *Journal of Chemical Theory and Computation* **2022**, *18*, 1–12.
- (39) Zills, F.; Schäfer, M. R.; Segreto, N.; Kästner, J.; Holm, C.; Tovey, S. Collaboration on Machine-Learned Potentials with IPSuite: A Modular Framework for Learning-on-the-Fly. *The Journal of Physical Chemistry B* **2024**,
- (40) Kellner, M.; Ceriotti, M. Uncertainty Quantification by Direct Propagation of Shallow Ensembles. *Machine Learning: Science and Technology* **2024**, *5*, 035006.
- (41) Kulichenko, M.; Barros, K.; Lubbers, N.; Li, Y. W.; Messerly, R.; Tretiak, S.; Smith, J. S.; Nebgen, B. Uncertainty-Driven Dynamics for Active Learning of Interatomic Potentials. *Nature Computational Science* **2023**, 1–10.
- (42) van der Oord, C.; Sachs, M.; Kovács, D. P.; Ortner, C.; Csányi, G. Hyperactive Learning for Data-Driven Interatomic Potentials. *npj Computational Materials* **2023**, *9*, 1–14.
- (43) Zaverkin, V.; Holzmüller, D.; Steinwart, I.; Kästner, J. Exploring Chemical and Conformational Spaces by Batch Mode Deep Active Learning. *Digit Discov* **2022**, *1*, 605–620.

- (44) Schoenholz, S. S.; Cubuk, E. D. JAX, M.D. A Framework for Differentiable Physics*. *Journal of Statistical Mechanics: Theory and Experiment* **2021**, *2021*, 124016.
- (45) Frank, J. T.; Unke, O. T.; Müller, K.-R.; Chmiela, S. A Euclidean transformer for fast and stable machine learned force fields. *Nature Communications* **2024**, *15*, 6539.
- (46) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **2007**, *98*, 146401.
- (47) Kocer, E.; Mason, J. K.; Erturk, H. A novel approach to describe chemical environments in high-dimensional neural network potentials. *The Journal of Chemical Physics* **2019**, *150*, 154102.
- (48) Seung, H. S.; Opper, M.; Sompolinsky, H. Query by committee. Proceedings of the fifth annual workshop on Computational learning theory. 1992; pp 287–294.
- (49) A. Peterson, A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics* **2017**, *19*, 10978–10985.
- (50) Xie, Y.; Vandermause, J.; Ramakers, S.; Protik, N. H.; Johansson, A.; Kozinsky, B. Uncertainty-aware molecular dynamics from Bayesian active learning for phase transformations and thermal transport in SiC. *npj Computational Materials* **2023**, *9*, 1–8.
- (51) Zhu, A.; Batzner, S.; Musaelian, A.; Kozinsky, B. Fast uncertainty estimates in deep learning interatomic potentials. *The Journal of Chemical Physics* **2023**, *158*, 164111.
- (52) Carrete, J.; Montes-Campos, H.; Wanzenböck, R.; Heid, E.; Madsen, G. K. H. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *The Journal of Chemical Physics* **2023**, *158*, 204801.
- (53) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On Calibration of Modern Neural

- Networks. Proceedings of the 34th International Conference on Machine Learning. 2017; pp 1321–1330.
- (54) Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. Proceedings of the 35th International Conference on Machine Learning. 2018; pp 2796–2804.
- (55) Gneiting, T.; and, A. E. R. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **2007**, *102*, 359–378.
- (56) Huber, P. J. Robust Estimation of a Location Parameter. *35*, 73–101, Publisher: Institute of Mathematical Statistics.
- (57) Rahimi, A.; Mensink, T.; Gupta, K.; Ajanthan, T.; Sminchisescu, C.; Hartley, R. Post-hoc Calibration of Neural Networks by g-Layers. 2022; <http://arxiv.org/abs/2006.12807>.
- (58) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.
- (59) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631–639.
- (60) Beveridge, D. L.; DiCapua, F. M. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annual Review of Biophysics* **1989**, *18*, 431–492.
- (61) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **1993**, *93*, 2395–2417.
- (62) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008**, *71*, 126601.

- (63) Kästner, J. Umbrella Sampling. *WIREs Computational Molecular Science* **2011**, *1*, 932–942.
- (64) Yoo, D.; Jung, J.; Jeong, W.; Han, S. Metadynamics sampling in atomic environment space for collecting training data for machine learning potentials. *npj Computational Materials* **2021**, *7*, 1–9.
- (65) Kirsch, A.; van Amersfoort, J.; Gal, Y. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. 2019; <http://arxiv.org/abs/1906.08158>.
- (66) Paszke, A. et al. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.
- (67) Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; <https://www.tensorflow.org/>.
- (68) de Buyl, P.; Colberg, P. H.; Höfling, F. H5MD: A structured, efficient, and portable file format for molecular data. *Computer Physics Communications* **2014**, *185*, 1546–1553.
- (69) Zills, F. Zincware/ZnH5MD: V0.3.6. 2024; <https://zenodo.org/records/13767043>.
- (70) Ziegler, J. F.; Biersack, J. P. In *Treatise on Heavy-Ion Science: Volume 6: Astrophysics, Chemistry, and Condensed Matter*; Bromley, D. A., Ed.; Springer US, 1985; pp 93–129.
- (71) Unke, O. T.; Maennel, H. E3x: E(3)-Equivariant Deep Learning Made Easy. 2024; <http://arxiv.org/abs/2401.07595>.
- (72) Batatia, I.; Batzner, S.; Kovács, D. P.; Musaelian, A.; Simm, G. N. C.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. The Design Space of E(3)-Equivariant Atom-Centered Interatomic Potentials. 2022; <http://arxiv.org/abs/2205.06643>.
- (73) DeepMind, et al. The DeepMind JAX Ecosystem. 2020; <http://github.com/google-deepmind>.

- (74) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
- (75) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2017; <https://arxiv.org/abs/1412.6980>.
- (76) Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-Aware Minimization for Efficiently Improving Generalization. 2021; <http://arxiv.org/abs/2010.01412>.
- (77) Pagliardini, M.; Ablin, P.; Grangier, D. The AdEMAMix Optimizer: Better, Faster, Older. 2024; <http://arxiv.org/abs/2409.03137>.
- (78) Huang, Y.; Cheng, Y.; Bapna, A.; Firat, O.; Chen, M. X.; Chen, D.; Lee, H.; Ngiam, J.; Le, Q. V.; Wu, Y.; Chen, Z. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc., 2019; pp 103–112.
- (79) Shazeer, N.; Cheng, Y.; Parmar, N.; Tran, D.; Vaswani, A.; Koanantakool, P.; Hawkins, P.; Lee, H.; Hong, M.; Young, C.; Sepassi, R.; Hechtman, B. MeshTensorFlow: deep learning for supercomputers. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018; pp 10435–10444.
- (80) Zills, F.; Schäfer, M. R.; Tovey, S.; Kästner, J.; Holm, C. ZnTrack – Data as Code. 2024; <https://arxiv.org/abs/2401.10603>.
- (81) Schran, C.; Brezina, K.; Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics* **2020**, *153*, 104105.
- (82) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials* **2020**, *6*, 1–11.
- (83) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M.

Uncertainty estimation for molecular dynamics and sampling. *The Journal of Chemical Physics* **2021**, *154*, 074102.

- (84) Chmiela, S.; Vassilev-Galindo, V.; Unke, O. T.; Kabylda, A.; Sauceda, H. E.; Tkatchenko, A.; Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances* **2023**, *9*, eadf0873.
- (85) Eckhoff, M.; Reiher, M. Lifelong Machine Learning Potentials. *Journal of Chemical Theory and Computation* **2023**, *19*, 3509–3525.
- (86) Qu, H.; Rahmani, H.; Xu, L.; Williams, B.; Liu, J. Recent Advances of Continual Learning in Computer Vision: An Overview. 2021; <http://arxiv.org/abs/2109.11369>.
- (87) Batatia, I. et al. A Foundation Model for Atomistic Materials Chemistry. 2023; <https://arxiv.org/abs/2401.00096>.
- (88) Frank, T.; Unke, O.; Müller, K.-R. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems. *Advances in Neural Information Processing Systems* **2022**, *35*, 29400–29413.
- (89) Kempen, L.; Nielsen, M.; Andersen, M. Breaking scaling relations with inverse catalysts: a machine learning exploration of trends in CO₂ hydrogenation energy barriers. 2025; <https://doi.org/10.48550/arXiv.2504.16493>.
- (90) Fuchs, P.; Chen, W.; Thaler, S.; Zavadlav, J. chemtrain-deploy: A parallel and scalable framework for machine learning potentials in million-atom MD simulations. 2025; <https://arxiv.org/abs/2506.04055>, Version Number: 1.
- (91) Jaffrelot Inizan, T.; Plé, T.; Adjoua, O.; Ren, P.; Gökcen, H.; Isayev, O.; Lagardère, L.; Piquemal, J.-P. Scalable hybrid deep neural networks/polarizable potentials biomolecular simulations including long-range effects. *Chemical Science* **2023**, *14*, 5438–5452, Publisher: Royal Society of Chemistry (RSC).

TOC Graphic

