Learning Atomistic Force Fields On-the-Fly with Bayesian Inference

Jonathan Vandermause,[1,2] Steven B. Torrisi,[1] Simon Batzner,[3] and Boris Kozinsky[2]

[1] *Department of Physics, Harvard University, Cambridge, MA 02138, USA*
[2] *John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA*
[3] *Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

(Dated: March 25, 2019)

Machine learning provides a path toward fast, accurate, and large-scale materials simulation, promising to combine the accuracy of *ab initio* methods with the computational efficiency of classical potentials. However, training current state-of-the-art models often requires hundreds of CPU hours and databases containing thousands of structures. We present an on-the-fly Bayesian inference scheme for automating and accelerating the construction of interatomic force fields. Gaussian process regression is coupled to a first principles DFT code to learn two- and three-body force fields on-the-fly with minimal trianing data. The resulting force field is easily extended to structures outside the training set and compares favorably to state-of-the-art classical and machine learned potentials.

Ab initio molecular dynamics is a powerful tool for accurately probing the dynamics of molecules and solids, but it is fundamentally limited by the cubic scaling of the most commonly used density functional theory (DFT) codes [1]. A common solution to this problem involves bypassing a quantum mechanical treatment of the electrons and instead directly modelling the Born-Oppenheimer potential energy surface of the ions. This is the approach taken when constructing classical interatomic potentials, which trade the accuracy of DFT and other first principles approaches for the speed and scalability of a local and analytic model, making possible the fully atomistic simulation of many thousands of atoms over nanosecond timescales. Classical potentials, however, have limited accuracy, flexibility, and transferability, and are inadequate in many settings.

Recent machine learning (ML) approaches to fitting interatomic potentials have been shown to approach first principles accuracy. However, most of these methods return only point estimates of the quantities of interest (typically energies, forces, and stess) rather than a predictive distribution reflecting model uncertainty. Without knowledge of the highest uncertainty training points, a laborious fitting procedure is required, in which thousands of reference structures are selected *ad hoc* from a database of first principles calculations. At test time, lack of predictive uncertainty makes it difficult to determine when the fitted model is out-of-sample, leading to unreliable results and making the model difficult to update in the presence of new data.

Here, we show that on-the-fly Bayesian inference can be used to both accelerate the training of a high-quality machine learned force field and flexibly adapt the model to out-of-sample structures. By coupling Gaussian process regression and density functional theory in a single molecular dynamics engine, it is shown that the number of DFT runs needed to train a high quality potential can be dramatically reduced from several thousand to a few dozen. By reducing the computational cost of both training and updating a high quality potential, our approach promises to extend ML modelling to a wide class of materials.

To reason about model uncertainty, we construct a Gaussian process model trained directly on *ab initio* forces. The model is trained on individual atomic environments rather than entire structures by expressing the total energy of the system as a sum over two- and three-body terms,

$$E = \sum_{ij} \varepsilon_{ij} + \sum_{ijk} \varepsilon_{ijk}, \tag{1}$$

where the sums range over all unique pairs and triplets of atoms containing at least one atom from the unfolded primary cell. In practice, the sums are truncated by considering local atom-centered environments surrounding each atom in the primary cell and neglecting contributions from atoms beyond a chosen cutoff distance from the central atom. The covariance between bond and triplet energies is set equal to a kernel defined directly over interatomic distances, from which a fully covariant and energy conserving force kernel follows immediately [2, 3]

[1] W. Kohn, Reviews of Modern Physics **71**, 1253 (1999).
[2] A. Glielmo, P. Sollich, and A. De Vita, Physical Review B **95**, 214302 (2017).
[3] A. Glielmo, C. Zeni, and A. De Vita, Physical Review B **97**, 184307 (2018).