The Journal of Open Source Software

SPARC: An Automated Workflow Toolkit for Accelerated Active Learning of Reactive Machine Learning Interatomic Potentials

Rahul Verma [1], **Nisarg Joshi** [1], **and Jim Pfaendtner** [1]

1 Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Evan Spotte-Smith

Reviewers:

- @mitwood
- @anjohan
- @PythonFZ

Submitted: 21 September 2025
Published: unpublished

Summary

Machine learning interatomic potentials (MLIPs) are rapidly becoming an essential modeling tool, as they allow atomistic simulations to achieve larger systems and longer time scales, while retaining the accuracy of ab-initio methods. However, constructing a reactive and transferable MLIP is still challenging, as it requires high quality training data including rare events, high energy intermediates and transition states.

Herein, we present SPARC (Smart Potential with Atomistic Rare events and Continuous learning), a modular Python workflow which automates the construction of reactive MLIPs that can generate chemically accurate potential energy surfaces (PES) for rare events. The idea for SPARC is to use an active learning (AL) protocol coupled with advanced sampling techniques that systematically identify new configurations and train ML models on-the-fly. The workflow consists of three main steps, which are executed iteratively in a loop until a reactive and stable MLIP is constructed. SPARC also includes utilities for visualizing configurations and plotting various properties to monitor the workflow across iterations.

Statement of need

While the AL protocol (Podryabinkin & Shapeev, 2017) (Miksch et al., 2021) has become a standard approach to refine ML models, it does not guarantee access to kinetically or thermodynamically rare events. Many learning schemes for MLIP invoke the use of higher temperature simulations to enhance phase space exploration. Even so, without additional enhanced sampling there is little to no possibility of systematically achieving reliable reactive MLIPs.

The motivation behind SPARC is to simplify the generation of high-quality training data and minimize the level of manual intervention required from the user. Although AL strategies have already proven effective for expanding training datasets, their integration with enhanced sampling techniques have typically been done in an ad hoc and largely manual manner. This makes it difficult to generalize these workflows beyond the original problem and limits reproducibility.

Several groups have coupled AL with enhanced sampling or pathway exploration techniques. Vitartas et al. (Vitartas et al., 2025) combined AL with well-tempered metadynamics to study organic reactions in both gas phase and solvent. Rivero et al. (Rivero et al., 2019) trained PhysNet architecture together with an AL at 1000K MD for Diels–Alder and hydrogen transfer reactions. Ang et al. (Ang et al., 2021) used SchNet combined with nudged elastic band calculations to explore solvent effects on pericyclic reactions. Parrinello and coworkers (Niu et al., 2020) coupled AL with variational enhanced sampling using DeePMD to investigate the phase diagram of gallium. Zhao et al. (Zhao et al., 2022) introduced a workflow combining umbrella sampling with AL to map out the solid-phase transition of GeSbTe. These studies highlight the promise of AL combined with enhanced sampling, but they remain highly tailored and driven by use case-specific implementations that can be difficult to generalize or reproduce.

SPARC automates this process by integrating the PLUMED library (Tribello et al., 2014) with AL protocol into a single modular workflow (see Figure 1). This allows user to use any advance sampling technique implemented in PLUMED to explore the configurational space and finding reactive configurations enabling scalable and reproducible MLIPs for generalized chemical environment.

Features and Implementation

SPARC is written in Python and can run on both local workstations and high-performance clusters. We utilized the Atomic Simulation Environment (ASE) (Larsen et al., 2017) library as the brain for SPARC which enables seamless integration of the MD engine, machine learning architecture, and electronic structure codes. To ensure portability, SPARC implements a file-based management system, and the output of each stage is written in a structured directory. The workflow iterations are stored in subdirectories (`iter_000000`, `iter_000001`,), with subfolders for reference first principle (FP) calculations (`00.dft`), MLIP training (`01.train`), and ML/MD simulations (`02.dpmd`). The SPARC workflow is controlled through a single YAML configuration file.

SPARC currently supports VASP (Hafner, 2008) and CP2K (Hutter et al., 2014) for electronic structure calculation, although this could be readily expanded in the future. For MLIP training we use DeePMD-kit architecture (Wang et al., 2018), with the ensemble model approach for query-by-committee (QbC) uncertainty estimation (Miksch et al., 2021). ML/MD simulations are run using ASE MD engine coupled together with both DeePMD and PLUMED calculator.

ML/MD output is stored in ASE trajectory formats as this enables the broader ecosystem of analysis tools that already support ASE compatible formats. Since SPARC manages all stages via ASE, the workfow requires a Python environment with necessary dependencies. This makes the workflow highly portable across computing environments and suitable for both exploratory studies and large-scale production run.

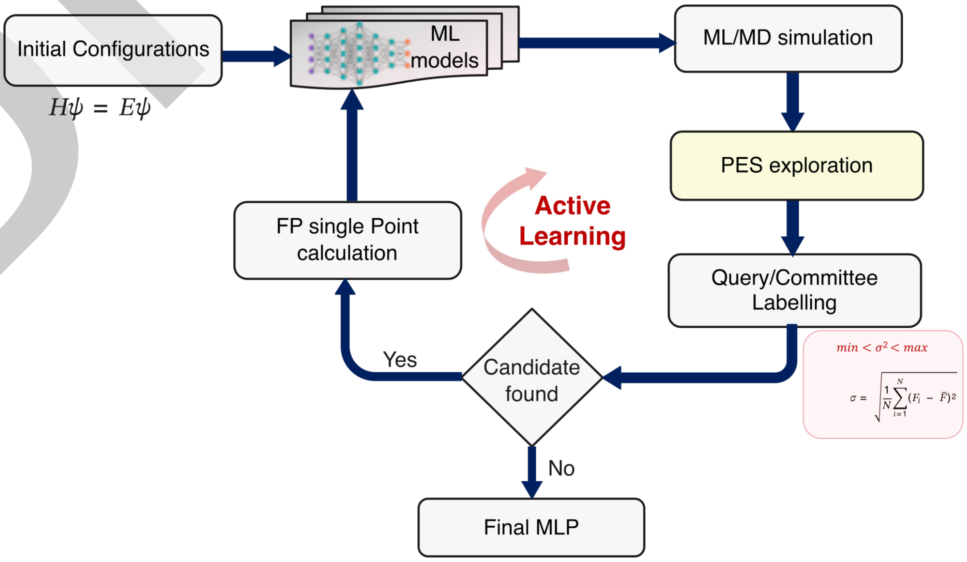

Figure 1: A schematic representation of AL cycle for training MLIPs. SPARC implements an additional block (PES exploration) to this cycle for systematic exploration of the configurational space.

Technical requirements and usage examples

After installation, a typical SPARC workflow can be launched with a single command:

```
sparc -i input.yaml
```

Illustrative Example: Potential Energy Scan

To illustrate the capabilities of SPARC, we applied the workflow to a simple ammonia borate (NH_3BH_3) molecule. The goal is to demonstrate how SPARC systematically expands training data and improves the accuracy of MLIPs through iterative learning. The initial training dataset only has 64 configurations, generated by scanning the B-N bond at the semiempirical level, followed by DFT single point calculations at PBE level with energy cutoff of 300 eV in VASP.

An ensemble of DeePMD models were trained and one of the ML models was used to run MD simulation for 5 ns with a timestep of 1 fs for each iteration. To ensure exploration beyond equilibrium structures, enhanced sampling was employed using parallel bias metadynamics (PbMetaD) (Pfaendtner & Bonomi, 2015) on SPRINT collective variables (Pietrucci & Andreoni, 2011). We obtain the uncertainty in forces by using an ensemble of trained models within QbC approach (Miksch et al., 2021) to flag configurations. Then configurations with standard deviation in the atomic forces between 0.05 to 0.5 eV/Å were automatically flagged and labeled with DFT. These structures were then added into the existing training dataset, after which the DeePMD models were retrained to get updated potential for the next cycle.

The effect of this iterative refinement is shown in Figure 2 which plots the maximum force deviation recorded in each cycle. In the initial iterations, deviations were large. As the workflow explores the chemical space and finds new configurations, these errors slowly decreased, indicating that the model was progressively learning the relevant physics.

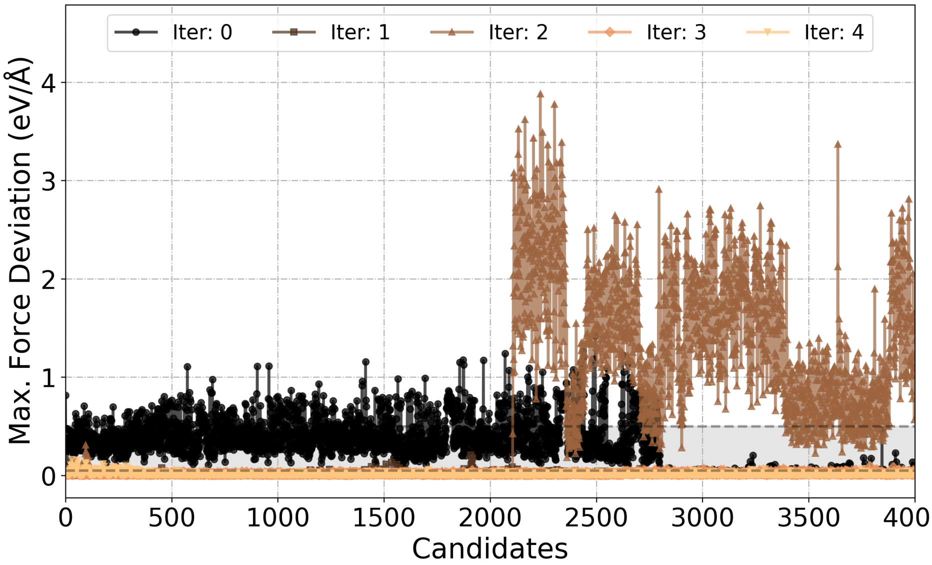

Figure 2: Force deviation across SPARC iterations. The shaded region marks the uncertainty thresholds for labeling.

By the fourth iteration, the error had converged to near-zero values, reflecting a stable and reactive MLIP. During exploration the model will be exposed to new configurations beyond training data which can result in very high forces, as observed in iteration 2.

We further assessed the reliability of trained MLIP under finite-temperature molecular dynamics. We performed enhanced sampling MD for both ab-initio and ML. In these simulations, the B-N bond distance was biased with metadynamics with a Gaussian width 0.05 Å and height 0.005 eV. The resulting free energy profile is shown in Figure 3.

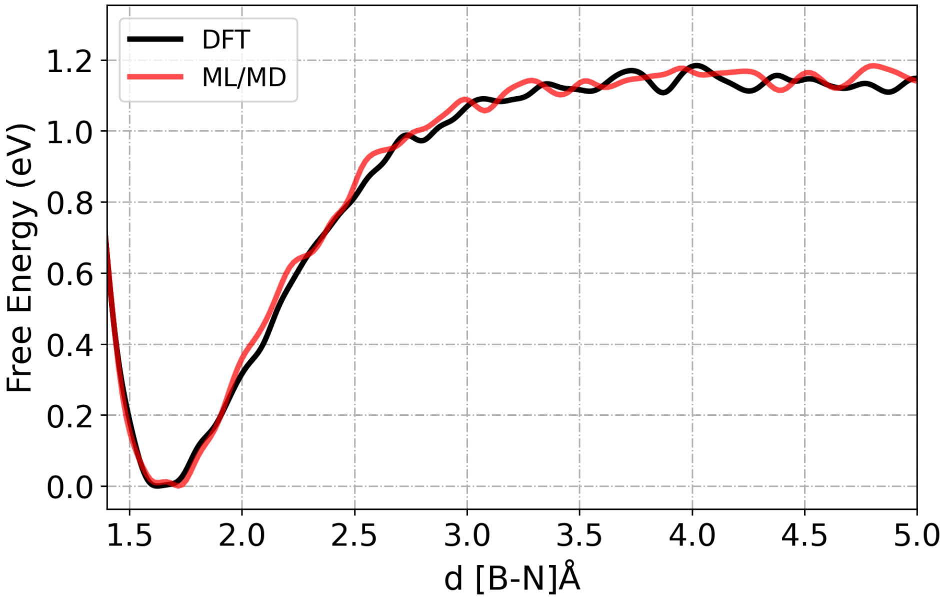

Figure 3: Free energy profile computed from both AIMD (black) and MLIP (red).

Here, MLIP was able to reproduce minima near 1.6 Å, with some minor discrepancies after 3.0 Å. The root mean square deviation between these curves was 0.04 eV, which is within chemical accuracy.

These results demonstrate MLIP trained with our workflow not only reproduces potential energy surface but also able to estimate quantitatively reliable predictions for computing free barriers and investigating chemical reactions.

Acknowledgements

This work was partially supported by NSF award EFRI-2132219 to RV and JP, and the HPC facility at the NC State University.

Related software

SPARC interface builds upon the following packages:

- ASE: atomic simulation setup, DFT calculators, and file handling.
- DeePMD-kit: MLIP training and deployment.
- PLUMED: enhanced sampling and CV biasing.
- VASP and CP2K: first-principles labeling.

References

Ang, S. J., Wang, W., Schwalbe-Koda, D., Axelrod, S., & Gómez-Bombarelli, R. (2021). Active learning accelerates ab initio molecular dynamics on reactive energy surfaces. *Chem*, *7*(3), 738–751. https://doi.org/10.1016/j.chempr.2020.12.009

Hafner, J. (2008). Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *Journal of Computational Chemistry*, *29*(13), 2044–2078. https://doi.org/10.1002/jcc.21057

Hutter, J., Iannuzzi, M., Schiffmann, F., & VandeVondele, J. (2014). CP2K: Atomistic simulations of condensed matter systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(1), 15–25. https://doi.org/10.1002/wcms.1159

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., & others. (2017). The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, *29*(27), 273002. https://doi.org/10.1088/1361-648X/aa680e

Miksch, A. M., Morawietz, T., Kästner, J., Urban, A., & Artrith, N. (2021). Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations. *Machine Learning: Science and Technology*, *2*(3), 031001. https://doi.org/10.1088/2632-2153/abfd96

Niu, H., Bonati, L., Piaggi, P. M., & Parrinello, M. (2020). Ab initio phase diagram and nucleation of gallium. *Nature Communications*, *11*(1), 2654. https://doi.org/10.1038/s41467-020-16372-9

Pfaendtner, J., & Bonomi, M. (2015). Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *Journal of Chemical Theory and Computation*, *11*(11), 5062–5067. https://doi.org/10.1021/acs.jctc.5b00846

Pietrucci, F., & Andreoni, W. (2011). Graph theory meets ab initio molecular dynamics: Atomic structures<? Format?> And transformations at the nanoscale. *Physical Review Letters*, *107*(8), 085504. https://doi.org/10.1103/PhysRevLett.107.085504

Podryabinkin, E. V., & Shapeev, A. V. (2017). Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, *140*, 171–180. https://doi.org/10.1016/j.commatsci.2017.08.031

Rivero, U., Unke, O. T., Meuwly, M., & Willitsch, S. (2019). Reactive atomistic simulations of diels-alder reactions: The importance of molecular rotations. *The Journal of Chemical Physics*, *151*(10). https://doi.org/10.1063/1.5114981

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., & Bussi, G. (2014). PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, *185*(2), 604–613. https://doi.org/10.1016/j.cpc.2013.09.018

Vitartas, V., Zhang, H., Juraskova, V., Johnston-Wood, T., & Duarte, F. (2025). Active learning meets metadynamics: Automated workflow for reactive machine learning potentials. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2024-twmlz

Wang, H., Zhang, L., Han, J., & E, W. (2018). DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, *228*, 178–184. https://doi.org/10.1016/j.cpc.2018.03.016

Zhao, Y., Sun, J., Yang, L., Zhai, D., Sun, L., & Deng, W. (2022). Umbrella sampling with machine learning potentials applied for solid phase transition of GeSbTe. *Chemical Physics Letters*, *803*, 139813. https://doi.org/10.1016/j.cplett.2022.139813