The Journal of Open Source Software

AniSOAP: Machine Learning Representations for Coarse-grained and Non-spherical Systems

Arthur Yan Lin [1], **Lucas Ortengren** [1], **Seonwoo Hwang**[1], **Yong-Cheol Cho** [1], **Jigyasa Nigam** [2], **and Rose K. Cersonsky** [1¶]

1 Department of Chemical and Biological Engineering, University of Wisconsin-Madison, USA **2** Research Laboratory of Electronics, Massachusetts Institute of Technology, USA **3** Department of Computer Science and Engineering, University of Wisconsin-Madison, USA ¶ Corresponding author

DOI: 10.21105/joss.07954

Software

- Review
- Repository
- Archive

Editor: Monica Bobra

Reviewers:

- @SamTov
- @DaniBodor

Submitted: 16 January 2025
Published: 10 July 2025

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

`AniSOAP` is a package that translates coarse-grained molecular configurations into tensorial representations that are ideal for supervised machine-learning models of thermodynamic quantities and unsupervised data-driven analyses. It generalizes several popular spherical representations for atomistic ML and aims to bridge the gap between two scientific communities: the machine-learning enabled atomistic simulation community, which leverages ML to accelerate and refine quantum modeling of complex interactions between spherical atoms, and the coarse-grained and colloid modeling community, which focuses on understanding emergent behavior of macroscopic particles with (plausibly) complex geometries. AniSOAP provides a common framework to tackle scientific questions at the intersection of these two fields.

Statement of need

Machine learning (ML) has greatly advanced atomistic molecular dynamics (MD), enabling both quick and quantum-accurate simulations and offering powerful tools for analyzing simulation results. Key to these advancements are the increasingly sophisticated strategies and software used to featurize atomistic environments that capture subtle differences between molecular configurations, either explicitly (Bartók et al., 2013; Behler, 2011; Drautz, 2019) or implicitly (Batatia et al., 2022; Batzner et al., 2022). These techniques have enabled supervised, semisupervised, and unsupervised studies across a wide variety of chemical spaces (Cersonsky et al., 2023; Cheng et al., 2019; De et al., 2016). However, these techniques are largely limited to atomistic resolution, and fall short in reliably describing coarse-grained entities (“particles’’ or groups of atoms) that have anisotropic geometries, where it is essential to resolve the orientation-dependence of their interactions with neighboring particles.

While many implementations construct spherical atomistic descriptors (e.g. DScribe (Himanen et al., 2020), librascal (*Librascal*, 2021; Musil et al., 2021), featomic (Fraux et al., 2025)), currently, there are no available packages for their anisotropic counterparts. In this software, we present the implementation of AniSOAP, an anisotropic generalization of the popular Smooth Overlap of Atomic Positions (SOAP) featurization (Bartók et al., 2013). SOAP, like other atomistic representations, offers a concise and numerically efficient parameterization of atomistic environments, incorporating correlations of the central atom with up to two of its neighbors. Along with several methods that refine its construction (Dusson et al., 2022; Nigam et al., 2020), it provides a framework to systematically build higher-order geometric and symmetrized “fingerprints’’ that can be used to model complex interaction potentials and extract machine-learning-enabled insights from data. AniSOAP extends this framework by allowing individual particles to be non-spherical. Hence, AniSOAP can be used as a

geometrically accurate, high body-order coarse-grained (CG) featurization of molecular and macromolecular systems. As AniSOAP retains full compatibility with SOAP, two representations can be used together to represent molecules at both atomistic and CG resolutions.

This is especially relevant as many systems cannot be simulated with all-atom resolution in reasonable times, despite high-performance capabilities of machine-learning interatomic potentials. Additionally, from a conceptual standpoint, we may not always want to analyze the behavior of *atoms*, but superatomic entities, such as functional groups. Most CG techniques still reduce macromolecules to a set of spherical beads. While often adequate for dilute simulations, this reduction to spherical beads significantly oversimplifies anisotropy, which is curcial in condensed systems (e.g., liquid crystals, glasses, molecular crystals). The flexibility of the AniSOAP representation coupled with learning algorithms can help address these challenges, as we can infer anisotropic coarse-graining from high-quality, first-principles data. This featurization could be easily plugged into other software, enabling mesoscopic simulations and enable data-driven insights to many chemical systems at different time and length scales.

The `AniSOAP` package enables the creation of AniSOAP feature vectors, which represent systems of ellipsoidal particles. Analogous to how SOAP or ACE create *atom*-centered representations, `AniSOAP` creates *particle*-centered representations, where a particle could be a single atom or a coarsened group of several atoms. Lin et al. (2024) provided demonstrations of specific use-cases of AniSOAP.

Implementation details

The AniSOAP package currently takes in as input a list of frames in the `Atomic Simulation Environment` package (Hjorth Larsen et al., 2017). Each frame contains the particles' positions, dimensions, and orientations. If using periodic boundary conditions, the frame also needs to contain the dimensions and orientations of the unit cell. Additional information about each frame can also be stored (e.g. the system energy) and used as a target for supervised ML.

With this information, one can construct an `EllipsoidalDensityProjection` object, whose main functionality is to calculate the expansion coefficients of an anisotropic density field in each frame via the `transform` method. Procedurally, calculating the expansion coefficients amounts to repeatedly and recursively computing high-order moments of an underlying multivariate Gaussian (Lin et al., 2024). For efficient computation, we have ported these highly-repeated calculations to Rust, a high-performance compiled language. The intermediate results utilize the `metatensor TensorMap` format (Fraux et al., 2024), which efficiently stores the AniSOAP featurizations and their associated metadata.

One can take Clebsch-Gordan products of these expansion coefficients to create higher body-order descriptors, and we optimize this step by caching intermediate results with a Least Recently Used (LRU) cache. This functionality is enabled by the `anisoap.metatensor_utils` module in AniSOAP, in particular, the `cg_combine` function.

As many users will be primarily interested in power-spectrum (i.e. 3-body) representations, we provide all the functionality required for these processes, and also provide the convenience method `power_spectrum` to calculate the 3-body descriptors of each frame. By default, this method returns the featurization as a $n_{\text{samples}} \times n_{\text{features}}$ `NumPy` array, which can be used as input into a machine learning algorithm. Alternatively, by setting the keyword argument `mean_over_samples=False`, this method returns a `metatensor TensorMap` object, which contains the power-spectrum representation for each atom in each frame as well as associated metadata. This is a much larger, unaggregated data object that requires more processing before it can be used in an ML algorithm. Examples of the various ways of creating AniSOAP representations can be found in the examples section of our documentation (Lin et al., 2025).

The library is thoroughly tested and documented, with unit-tests to test basic functionality,

integration-tests to ensure that AniSOAP vectors are calculated correctly, and caching and speed tests to ensure that our aforementioned optimizations yield faster code. These tests are integrated into a Github Continuous Integration (CI), and we ensure that future features should necessistate additional tests and should pass existing ones.

Conclusion and future developments

AniSOAP is a powerful featurization that can be used for supervised and unsupervised analyses of molecular systems. AniSOAP is under active development and we envision it being used in a wide variety of contexts. Our main future development goals involve using AniSOAP as the underlying representation for machine-learned anisotropic potentials, and to understand how the relationship between AniSOAP and its all-atom counterpart SOAP fits into the broad theory of bottom-up coarse-graining. We hope that accomplishing these goals can enable fast, accurate, and interpretable macromolecular or colloidal simulations.

Acknowledgements

This project was funded by the Wisconsin Alumni Research Fund (R.K.C.), NSF through the University of Wisconsin Materials Research Science and Engineering Center (Grant No. DMR-2309000, A.L.), the European Research Council (ERC) under the research and innovation program (Grant Agreement No. 101001890-FIAMMA, J.N.), and the MIT Postdoc Fellowship for Excellence in Engineering (PFPFEE, J.N.).

We extend our un-ending gratitude to Guillaume Fraux and the developers of featomic for fielding our many questions during the implementation and validation of AniSOAP, and Kevin Kazuki Huguenin-Dumittan for building the first iteration of AniSOAP.

References

Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. *Physical Review B*, *87*(18), 184115. https://doi.org/10.1103/PhysRevB.87.184115

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., & Csányi, G. (2022). MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. ISBN: 9781713871088

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., & Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-29939-5

Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, *134*(7), 074106. https://doi.org/10.1063/1.3553717

Cersonsky, R. K., Pakhnova, M., Engel, E. A., & Ceriotti, M. (2023). A data-driven interpretation of the stability of organic molecular crystals. *Chemical Science*, *14*(5), 1272–1285. https://doi.org/10.1039/D2SC06198H

Cheng, B., Engel, E. A., Behler, J., Dellago, C., & Ceriotti, M. (2019). Ab initio thermodynamics of liquid and solid water. *Proceedings of the National Academy of Sciences*, *116*(4), 1110–1115. https://doi.org/10.1073/pnas.1815117116

De, S., Bartók, A. P., Csányi, G., & Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, *18*(20),

13754–13769. https://doi.org/10.1039/C6CP00415F

Drautz, R. (2019). Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, *99*(1), 014104. https://doi.org/10.1103/PhysRevB.99.014104

Dusson, G., Bachmayr, M., Cs'anyi, G., Drautz, R., Etter, S., Der Oord, C. van, & Ortner, C. (2022). Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, *454*, 110946. https://doi.org/10.1016/j.jcp.2022.110946

Fraux, G., Loche, P., Kliavinek, S., Kazuki Huguenin-Dummitan, K., Tisi, D., & Goscinski, A. (2025). *Featomic*. https://github.com/metatensor/featomic

Fraux, G., Tisi, D., Loche, P., Abbott, J. W., Nigam, J., & Mahmoud, C. B. (2024). *Metatensor* [Documentation]. https://docs.metatensor.org/latest/index.html

Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., Rinke, P., & Foster, A. S. (2020). DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, *247*, 106949. https://doi.org/10.1016/j.cpc.2019.106949

Hjorth Larsen, A., Jørgen Mortensen, J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Bjerre Jensen, P., Kermode, J., Kitchin, J. R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., … Jacobsen, K. W. (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

Librascal. (2021). https://doi.org/10.5281/zenodo.4526062

Lin, A., Huguenin-Dumittan, K. K., Cho, Y.-C., Nigam, J., & Cersonsky, R. K. (2024). Expanding density-correlation machine learning representations for anisotropic coarse-grained particles. *The Journal of Chemical Physics*, *161*(7), 074112. https://doi.org/10.1063/5.0210910

Lin, A., Ortengren, L., Hwang, S., Cho, Y.-C., Nigam, J., & Cersonsky, R. K. (2025). *AniSOAP* [Documentation]. https://anisoap.readthedocs.io/en/latest/

Musil, F., Veit, M., Goscinski, A., Fraux, G., Willatt, M. J., Stricker, M., Junge, T., & Ceriotti, M. (2021). Efficient implementation of atom-density representations. *The Journal of Chemical Physics*, *154*(11). https://doi.org/10.1063/5.0044689

Nigam, J., Pozdnyakov, S., & Ceriotti, M. (2020). Recursive evaluation and iterative contraction of n-body equivariant features. *The Journal of Chemical Physics*, *153*(12). https://doi.org/10.1063/5.0021116