The Journal of Open Source Software

Xponge: A Python package to perform pre- and post-processing of molecular simulations

Yijie Xia [1] **and Yi Qin Gao**[1¶]

1 College of Chemistry and Molecular Engineering, Peking University, China ¶ Corresponding author

DOI: 10.21105/joss.04467

Software

- Review
- Repository
- Archive

Editor: Richard Gowers

Reviewers:

- @ojeda-e
- @hmacdope

Submitted: 22 May 2022
Published: 22 September 2022

Summary

Xponge is a lightweight and easy to customize Python package to perform pre- and post-processing of molecular simulations. It is mainly designed for the MD program SPONGE (Huang et al., 2022), but it can also process common format files and therefore it should also be useful for other simulation packages such as GROMACS (Abraham et al., 2015) and LAMMPS (Thompson et al., 2022). Xponge includes three major categories of functionality, namely, the simulation system construction, simulation data transformation and analysis, and automated workflows for complex simulations. For the construction of simulation systems, Xponge can generate 3-dimensional molecular structures or read structures downloaded from online databases such as RCSB (Berman et al., 2000) and PubChem (Sayers et al., 2021), and perform force field parameterization. The current force fields supported by Xponge include CHARMM27 (MacKerell et al., 1998; Mackerell Jr. et al., 2004), ff14SB (Maier et al., 2015), ff19SB (Tian et al., 2020) for proteins, lipid14 (Dickson et al., 2014) and lipid17 for lipids, GAFF (Wang et al., 2004) for small organic molecules and SPC/E (Berendsen et al., 1987), TIP3P (Jorgensen et al., 1983), TIP4P-Ew (Horn et al., 2004), OPC (Izadi et al., 2014) for water. Simulation data transformation is the process of changing the format, structure, or values of the simulation data such as molecular dynamics trajectories and coordinates from any frame of a simulation. For both simulation data transformation and analysis, Xponge is combined with the Python package MDAnalysis (Gowers et al., 2016; Michaud-Agrawal et al., 2011) for processing. Xponge has an integrated workflow for free energy perturbation calculations with dual topology construction, facilitating calculations on hydration and binding free energies. At the same time, Xponge itself is highly modular and easily customizable, enabling simple extensions that mimic existing modules to develop one's own force fields, data analytics, and workflows.

Statement of need

Molecular simulation is becoming an important and useful tool in many different research areas. For example, the computational simulation of proteins and small organic molecules in water is now widely used in drug design (Yang et al., 2016). In MD simulations, it is important to construct and parameterize the simulation system and to process and analyze the simulation results. Because of the complexity and variety of systems to be simulated, it is often necessary to develop new and different force fields, approximations, algorithms, or analytical methods to obtain and visualize results. There exist a variety of pre-processing tools for molecular simulations, such as offline LEaP (Li & Cerutti, 2022), pdb2gmx (Lindahl, 2022) and psfgen (Joao V. Ribeiro & Phillips, 2022), and online LigParGen (Leela S. Dodda & Jorgensen, 2022), CGenFF (Vanommeslaeghe et al., 2012) and ATB (Malde et al., 2011; Stroet et al., 2018), but they are designed for a particular form of force field and can be difficult to customize. Xponge is designed to be highly modular and intended to be developer-friendly. When a developer wants to develop his own force field, it is hard to modify the online tools and the

offline tools usually only have parameter interfaces. While Xponge provides an interface that allows users to specify their forcefields themselves without modifying the source code. At the same time, Xponge includes a number of post-processing capabilities and combines specific complex pre-processing - simulation - post-processing into one single workflow. Meanwhile, since machine learning force field for molecular simulation represents one of the current research frontiers (Poltavsky & Tkatchenko, 2021; Unke et al., 2021), and Python is now the top programming language for machine learning (Elliott, 2019). Xponge is written by python and has interfaces to call the machine learning framework MindSpore (MindSpore, 2022), thus Xponge can be better merged into the computational graph (TutorialsPoint, 2018). Such a setup fits well with molecular simulation softwares such as SPONGE (Huang et al., 2022) which incorporates machine learning methods.

Availability

Xponge is freely available and open source under the Apache License 2.0 (Apache-2.0). You can download the package and access the online documentations on gitee or github.

Acknowledgements

The authors thank the National Key R&D Program of China (2017YFA0204702), the National Natural Science Foundation of China (21821004, 21927901, 92053202 and 22050003) for financial support.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Berendsen, H. J. C., Grigera, J. R., & Straatsma, T. P. (1987). The missing term in effective pair potentials. *The Journal of Physical Chemistry*, *91*(24), 6269–6271. https://doi.org/10.1021/j100308a038

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Dickson, C. J., Madej, B. D., Skjevik, A. A., Betz, R. M., Teigen, K., Gould, I. R., & Walker, R. C. (2014). Lipid14: The amber lipid force field. *Journal of Chemical Theory and Computation*, *10*(2), 865–879. https://doi.org/10.1021/ct4010307

Elliott, T. (2019). *The state of the octoverse: Machine learning*. https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/

Gowers, Richard J., Linke, Max, Barnoud, Jonathan, Reddy, Tyler J. E., Melo, Manuel N., Seyler, Sean L., Domański, Jan, Dotson, David L., Buchoux, Sébastien, Kenney, Ian M., & Beckstein, Oliver. (2016). MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Sebastian Benthall & Scott Rostrup (Eds.), *Proceedings of the 15th Python in Science Conference* (pp. 98–105). https://doi.org/10.25080/Majora-629e541a-00e

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., & Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-ew. *The Journal of Chemical Physics*, *120*(20), 9665–9678. https://doi.org/10.1063/1.1683075

Huang, Y.-P., Xia, Y., Yang, L., Wei, J., Yang, Y. I., & Gao, Y. Q. (2022). SPONGE: A GPU-accelerated molecular dynamics package with enhanced sampling and AI-driven algorithms. *Chinese Journal of Chemistry*, *40*(1), 160–168. https://doi.org/10.1002/cjoc.202100456

Izadi, S., Anandakrishnan, R., & Onufriev, A. V. (2014). Building water models: A different approach. *The Journal of Physical Chemistry Letters*, *5*(21), 3863–3871. https://doi.org/10.1021/jz501780a

Joao V. Ribeiro, J. S., Brian Radak, & Phillips, J. (2022). *VMD psfgen plugin, version 2.0*. https://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*(2), 926–935. https://doi.org/10.1063/1.445869

Leela S. Dodda, J. T.-R., Israel Cabeza de Vaca, & Jorgensen, W. L. (2022). *LigParGen server*. http://zarbi.chem.yale.edu/ligpargen/

Li, P., & Cerutti, D. (2022). *Fundamentals of LEaP*. https://ambermd.org/tutorials/pengfei/index.php

Lindahl, P. B. B. H. E. (2022). *Gmx pdb2gmx*. https://doi.org/10.5281/zenodo.7037337

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., … Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. https://doi.org/10.1021/jp973084f

Mackerell Jr., A. D., Feig, M., & Brooks III, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, *25*(11), 1400–1415. https://doi.org/10.1002/jcc.20065

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., Oostenbrink, C., & Mark, A. E. (2011). An automated force field topology builder (ATB) and repository: Version 1.0. *Journal of Chemical Theory and Computation*, *7*(12), 4026–4037. https://doi.org/10.1021/ct200196m

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, *32*(10), 2319–2327. https://doi.org/10.1002/jcc.21787

MindSpore. (2022). *MindSpore: An open AI framwork*. http://www.mindspore.cn

Poltavsky, I., & Tkatchenko, A. (2021). Machine learning force fields: Recent advances and remaining challenges. *The Journal of Physical Chemistry Letters*, *12*(28), 6551–6564. https://doi.org/10.1021/acs.jpclett.1c01204

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Stroet, M., Caron, B., Visscher, K. M., Geerke, D. P., Malde, A. K., & Mark, A. E. (2018). Automated topology builder version 3.0: Prediction of solvation free enthalpies in water

and hexane. *Journal of Chemical Theory and Computation*, *14*(11), 5834–5845. https://doi.org/10.1021/acs.jctc.8b00768

Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C., & Plimpton, S. J. (2022). LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, *271*, 108171. https://doi.org/10.1016/j.cpc.2021.108171

Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., & Simmerling, C. (2020). ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of Chemical Theory and Computation*, *16*(1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591

TutorialsPoint. (2018). *Python deep learning: Computational graphs*. https://www.tutorialspoint.com/python_deep_learning/python_deep_learning_computational_graphs.htm

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., & Müller, K.-R. (2021). Machine learning force fields. *Chemical Reviews*, *121*(16), 10142–10186. https://doi.org/10.1021/acs.chemrev.0c01111

Vanommeslaeghe, K., Raman, E. P., & MacKerell, A. D. (2012). Automation of the CHARMM general force field (CGenFF) II: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling*, *52*(12), 3155–3168. https://doi.org/10.1021/ci3003649

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, *25*(9), 1157–1174. https://doi.org/10.1002/jcc.20035

Yang, L., Zhang, J., Che, X., & Gao, Y. (2016). Simulation studies of protein and small molecule interactions and reaction. In *Methods in Enzymology* (Vol. 578). https://doi.org/10.1016/bs.mie.2016.05.031