Newcastle University

Open Research Case Study

Supporting an open science, open source, open data approach to force field design and drug discovery with the Open Force Field Initiative

Josh Horton

Research Associate. School of Natural and Environmental Sciences.

Introduction and Research Context:

The efficiency of drug discovery has been falling for decades such that, for each new molecule that reaches the consumer, estimated research and development costs are over $2 billion. Computational models capable of describing the dynamics and interactions of biological molecules at the atomistic scale allow researchers to study the molecular mechanisms of diseases and design new therapeutics. A crucial enabling technology in this research area is the molecular mechanics force field, which is a parameterised mathematical model describing the interactions between atoms. The parameters behind the model, until recently, have been curated over many decades by a handful of research groups. The provenance of many of the parameters is unknown or forgotten.

To combat this issue, the Open Force Field (OpenFF) Initiative (https://openforcefield.org/) was set up as an academic-industry collaboration to bring an open and collaborative approach to better force fields. I joined OpenFF in 2019, and continue collaborating with them in my current role as a research associate in Newcastle Chemistry. Chemical space is vast and designing a simple force field that covers all use cases is challenging. The partnership between OpenFF and industry potentially gives academics access to valuable information about use cases, but this presents a challenge as much of pharmaceutical work is proprietary. Hence, in collaboration with ten industrial partners, we helped coordinate a large-scale, open benchmarking of force fields on pharmaceutically relevant molecules from each company's internal programs [1]. This provided a unique opportunity for industry partners to engage with open science practices, and one another, to benefit the computational chemistry community.

Open research practices

To facilitate fair and reproducible benchmarking across the partner sites an open-source benchmarking software suite was developed (https://github.com/openforcefield/openff-benchmark) to execute the redefined workflow on the proprietary in-house data. The open-source software allows anyone from the community to repeat or improve the benchmark by contributing to the software, enabling development into a community best practice as it evolves.

To ensure that future generations of force fields show improved accuracy in pharmaceutically-relevant areas of chemical space, we also collated an open dataset of drug-like molecules from industry past projects. This generated a high-quality dataset with over 77,000 data points for around 10,000 unique molecules. The dataset is now used to benchmark all OpenFF small molecule force fields [2], and is freely available from an open database (QCArchive).

Furthermore, users often want to improve the accuracy of a force field for specific areas of chemical space without revealing proprietary information. So we developed an open-source software package, named BespokeFit (https://github.com/openforcefield/openff-bespokefit), which generates molecule-specific parameters on-the-fly using an automated workflow [3].

To enable maximum impact and reach of the work, the publications (CC-BY and chemrxiv preprint service), software (github, MIT licence) and data (Zenodo) were released under open licence. To aid the adoption and dissemination of the software, comprehensive documentation is provided with detailed examples and explanations of the algorithms used (e.g. https://docs.openforcefield.org/projects/bespokefit/en/latest/).

Benefits, Challenges and Lessons Learned

Providing an open-source benchmarking suite and force field fitting software allows industry partners to improve force field accuracy without disclosing proprietary information. Moreover,due to the permissive MIT license, the bespoke fitting workflow has been integrated into commercial computer-aided drug design software (https://www.cresset-group.com/about/news/ff-md/), thereby impacting live drug discovery campaigns.

The open-source data curation tooling (https://github.com/openforcefield/openff-qcsubmit) from this project has made it trivial to generate large-scale datasets and led to the creation of SPICE [4], one of the most comprehensive open machine learning quantum chemistry datasets. This has benefited Newcastle University through a collaboration with the University of Cambridge whereby a state-of-the-art general organic machine learning force field (MACE-OFF23) has been trained on this dataset [5].

Developing open-source software which follows best practices (commented code, documentation, unit tests, continuous integration and deployment packages) comes with additional time costs. However, following these principles pays dividends with increased adoption due to ease of installation and confidence in the ability of a well-tested package to provide reproducible results. To follow these best practices, we used the MolSSI cookiecutter (https://github.com/MolSSI/cookiecutter-cms), which provides a template for a software repository that follows these guidelines.

Collecting and combining research outputs from different practitioners can be difficult with different standards for units and formats, hence we designed an output data structure for the benchmarking software that reported the results in a consistent programmatically readable format. Reproducibility is a long-standing challenge in science that can be addressed by publishing software, workflows, tutorials and datasets openly. This also empowers the community to build upon the work quickly, demonstrating greater impact and unlocking future collaborations. This is exemplified by the systematic improvements in OpenFF force fields that address problems identified in the collaborative benchmarking project.

Conclusion

By engaging with industry in open science projects we have laid a foundation to shape a culture of open research in our field, which will lead to improved physical models that, in turn, will directly impact drug discovery. I am grateful to collaborators at the Open Force Field Initiative for their leadership and support!

References

1. D'Amore, L., Hahn, D.F., Dotson, D.L., Horton, J.T., Anwar, J., Craig, I., Fox, T., Gobbi, A., Lakkaraju, S.K., Lucas, X. and Meier, K., 2022. Collaborative assessment of molecular geometries and energies from the Open Force Field. *Journal of chemical information and modeling*, *62*(23), pp.6094-6104.

2. Boothroyd, S., Behara, P.K., Madin, O.C., Hahn, D.F., Jang, H., Gapsys, V., Wagner, J.R., Horton, J.T., Dotson, D.L., Thompson, M.W. and Maat, J., 2023. Development and benchmarking of open force field 2.0. 0: the Sage small molecule force field. *Journal of chemical theory and computation*, *19*(11), pp.3251-3275.

3. Horton, J.T., Boothroyd, S., Wagner, J., Mitchell, J.A., Gokey, T., Dotson, D.L., Behara, P.K., Ramaswamy, V.K., Mackey, M., Chodera, J.D. and Anwar, J., 2022. Open force field BespokeFit: automating bespoke torsion parametrization at scale. *Journal of chemical information and modeling*, *62*(22), pp.5622-5633.

4. Eastman, P., Behara, P.K., Dotson, D.L., Galvelis, R., Herr, J.E., Horton, J.T., Mao, Y., Chodera, J.D., Pritchard, B.P., Wang, Y. and De Fabritiis, G., 2023. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, *10*(1), p.11.

5. Kovács, D.P., Moore, J.H., Browning, N.J., Batatia, I., Horton, J.T., Kapil, V., Magdău, I.B., Cole, D.J. and Csányi, G., 2023. Mace-off23: Transferable machine learning force fields for organic molecules. *arXiv preprint arXiv:2312.15211*.

