The Journal of Open Source Software

kallisto: A command-line interface to simplify computational modelling and the generation of atomic features

Eike Caldeweyher[*1]

1 Data Science and Modelling, Pharmaceutical Science, R&D, AstraZeneca, Gothenburg, Sweden

DOI: 10.21105/joss.03050

Software

- Review
- Repository
- Archive

Editor: Richard Gowers

Reviewers:

- @rmeli
- @Sulstice

Submitted: 05 February 2021
Published: 15 April 2021

Statement of Need

Machine learning (ML) has recently become very popular within pharmaceutical industry (Roy et al., 2015; Sprous et al., 2010). Tasks as, e.g., building predictive models, performing virtual screening, or predicting compound activities are potential use cases for such ML applications (Li et al., 2021; Simm et al., 2018). Traditionally, ML models often rely on the quantitative structure-activity relationship (QSAR) that has been popularized by medicinal chemists and statisticians to relate bioactivities to specific functional group manipulations (Dudek et al., 2006; Verma et al., 2010). This QSAR approach decreases the dimensionality of the underlying problem and projects the molecular structure into a space spanned by the physicochemical features. While early approaches relied more on linear regression, modern approaches combine such features with non-linear ML algorithms.

Chemoinformatic packages like RDKit (Landrum & others, 2006) enable the fast calculation of atomic/molecular features based on structural information like the molecular graph, while recently an extended Hueckel package has been added as well (Landrum, 2019). However, frequently we want to go beyond a structure-only approach thus incorporating electronic structure effects as obtained, e.g., by a (higher-level) quantum mechanical (QM) treatment. The calculation of QM-based features relies often on well-established quantum chemistry methods like Kohn-Sham density functional theory (DFT) that is currently the workhorse of computational chemistry (Kohn, 1999; Parr, 1980). However, generating the feature space by DFT is computationally demanding and can become the computational bottleneck especially when aiming for high-throughput experiments with several hundred to thousands of molecules.

Since there exists a critical need for an efficient yet accurate featurizer, we developed the `kallisto` command-line interface that is able to calculate QM-based atomic features for atoms and molecules efficiently (whole periodic table up to Radon). The features are either interpolating high-level references (e.g., static/dynamic polarizabilities with time-dependent DFT data) or are parametrized (Caldeweyher et al., 2019) to reproduce QM references (e.g., DFT Hirshfeld (Hirshfeld, 1977) atomic partial charges). Molecular geometries need to have an `xmol` or a `Turbomole` like format to be processed by `kallisto`. Besides, we implemented several computational modelling helpers to simplify the development of high-throughput procedures. Some of those modelling helpers depend on the open-source xtb tight-binding scheme that has been developed by Stefan Grimme and co-worker (Bannwarth et al., 2020). The `kallisto` software depends on the scientific libraries Numpy (Harris et al., 2020) and SciPy (Virtanen et al., 2020). The online documentation covers all high-level functionalizations of this software mostly in terms of copy-paste recipes. Furthermore, we cover bits of the underlying theory and compare to experimental data as well as to other modern deep learning models.

*Corresponding author

Atomic and Molecular Features

The following atomic and molecular features are available for all atoms up to Radon

- Coordination numbers (Caldeweyher et al., 2019; Grimme et al., 2010)
- Proximity shells
- Environment-dependent electronegativity equilibration partial charges (Caldeweyher et al., 2019)
- Environment- and charge-dependent dynamic polarizabilities (Caldeweyher et al., 2019; Grimme et al., 2010)
- Environment- and charge-dependent van-der-Waals radii (Fedorov et al., 2018; Mantina et al., 2009; Rahm et al., 2017)
- Sterimol descriptors (L, Bmin, Bmax) (Brethome et al., 2019)

Modelling Helpers

The following modelling helper are implemented

- Breadth first sorting
- Root mean squared deviation (quaternions) (Coutsias et al., 2004)
- Substructure identifier
- Substructure exchanger

Acknowledgements

EC acknowledges contributions from Philipp Pracht (`@pprcht`) and thanks Kjell Jorner (`@kjelljorner`) for sharing his Sterimol algorithm.

References

Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., & Grimme, S. (2020). Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.*, e01493. https://doi.org/10.1002/wcms.1493

Brethome, A. V., Fletcher, S. P., & Paton, R. S. (2019). Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catalysis*, *9*(3), 2313–2323. https://doi.org/10.1021/acscatal.8b04043

Caldeweyher, E., Ehlert, S., Hansen, A., Neugebauer, H., Spicher, S., Bannwarth, C., & Grimme, S. (2019). A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.*, *150*(15), 154122. https://doi.org/10.1063/1.5090222

Coutsias, E. A., Seok, C., & Dill, K. A. (2004). Using quaternions to calculate RMSD. *J. Comput. Chem.*, *25*(15), 1849–1857. https://doi.org/10.1002/jcc.20110

Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screen.*, *9*(3), 213–228. https://doi.org/10.2174/138620706776055539

Fedorov, D. V., Sadhukhan, M., Stöhr, M., & Tkatchenko, A. (2018). Quantum-mechanical relation between atomic dipole polarizability and the van der Waals radius. *Phys, Rev. Lett.*, *121*(18), 183401. https://doi.org/10.1103/PhysRevLett.121.183401

Grimme, S., Antony, J., Ehrlich, S., & Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, *132*(15), 154104. https://doi.org/10.1063/1.3382344

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., & others. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hirshfeld, F. L. (1977). Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta*, *44*(2), 129–138. https://doi.org/10.1007/bf00549096

Kohn, W. (1999). Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Rev. Mod. Phys.*, *71*(5), 1253. https://doi.org/10.1103/revmodphys.71.1253

Landrum, G. (2019). YAeHMOP: Yet Another extended Hueckel Molecular Orbital Package. In *GitHub repository*. GitHub. https://github.com/greglandrum/yaehmop

Landrum, G., & others. (2006). *RDKit: Open-source cheminformatics*. https://www.rdkit.org/

Li, H., Sze, K.-H., Lu, G., & Ballester, P. J. (2021). Machine-learning scoring functions for structure-based virtual screening. *WIREs Comput. Mol. Sci.*, *11*(1), e1478. https://doi.org/10.1002/wcms.1478

Mantina, M., Chamberlin, A. C., Valero, R., Cramer, C. J., & Truhlar, D. G. (2009). Consistent van der Waals radii for the whole main group. *J. Phys. Chem. A*, *113*(19), 5806–5812. https://doi.org/10.1021/jp8111556

Parr, R. G. (1980). *Density Functional Theory of Atoms and Molecules* (pp. 5–15). Springer-Verlag GmbH, Heidelberg. https://doi.org/10.1007/978-94-009-9027-2_2

Rahm, M., Hoffmann, R., & Ashcroft, N. W. (2017). Corrigendum: Atomic and Ionic Radii of Elements 1–96. *Chem. Eur. J.*, *23*(16), 4017–4017. https://doi.org/10.1002/chem.201700610

Roy, K., Kar, S., & Das, R. N. (2015). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Elsevier Inc. https://doi.org/10.1016/c2014-0-00286-9

Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J. K., Gustin, E., Chupakhin, V., Chong, Y. T., Vialard, J., Buijnsters, P., & others. (2018). Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.*, *25*(5), 611–618. https://doi.org/10.1016/j.chembiol.2018.01.015

Sprous, D. G., Palmer, R. K., Swanson, J. T., & Lawless, M. (2010). QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr. Top. Med. Chem.*, *10*(6), 619–637. https://doi.org/10.2174/156802610791111506

Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design-a review. *Curr. Top. Med. Chem.*, *10*(1), 95–115. https://doi.org/10.2174/156802610790232260

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., & others. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2