The Journal of Open Source Software

OpenFEPOPS: A Python implementation of the FEPOPS molecular similarity technique

Yan-Kai Chen [1], **Douglas R. Houston** [1], **Manfred Auer** [1,2], **and Steven Shave** [1¶]

1 School of Biological Sciences, University of Edinburgh, The King's Buildings, Max Born Crescent, CH Waddington Building, Edinburgh, EH9 3BF, United Kingdom. **2** Xenobe Research Institute, P. O. Box 3052, San Diego, California, 92163, United States. ¶ Corresponding author

DOI: 10.21105/joss.05763

Software

- Review
- Repository
- Archive

Editor: Richard Gowers

Reviewers:

- @hannahbaumann
- @exs-cbouy

Submitted: 06 August 2023
Published: 09 November 2023

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

Summary

OpenFEPOPS is an open-source Python implementation of the FEature POint PharmacophoreS (FEPOPS) molecular similarity technique (Jenkins et al., 2004; Jenkins, 2013; Nettles et al., 2007) enabling descriptor generation, comparison, and ranking of molecules in virtual screening campaigns. Ligand based virtual screening (Ripphausen et al., 2011) is a fundamental approach undertaken to expand hit series or perform scaffold hopping whereby new chemistries and synthetic routes are made available in efforts to remove undesirable molecular properties and discover better starting points in the early stages of drug discovery (Hughes et al., 2011). Typically, these techniques query hit molecules against proprietary, in-house, or publicly available repositories of small molecules in the hope of finding close matches which will display similar activities to the query based on the molecular similarity principle which states that similar molecules should have similar properties and make similar interactions (Cortés-Ciriano et al., 2020). Often batteries of these similarity measures are used in parallel, helping to score molecules from many different subjective viewpoints and measures of similarity (Baber et al., 2006). The central idea behind FEPOPS is reducing the complexity of molecules by merging of local atomic environments and atom properties into 'feature points'. This compressed feature point representation has been used to great effect as noted in literature, helping researchers identify active and potentially therapeutically valuable small molecules. By default, OpenFEPOPS uses literature reported parameters which show good performance in retrieval of active lead- and drug-like small molecules within virtual screening campaigns, with feature points capturing charge, lipophilicity, and hydrogen bond acceptor and donor status. When run with default parameters, OpenFepops compactly represents molecules using seven sets of four feature points, with each feature point encoded into 22 numeric values, resulting in a compact representation of 616 bytes per molecule. By extension, this allows the indexing of a compound archive containing 1 million small molecules using 587.5 MB of data. Whilst more compact representations are readily available, the FEPOPS technique strives to capture tautomer and conformer information, first through enumeration and then through diversity driven selection of representative FEPOPS descriptors to capture the diverse states that a molecule may adopt.

Statement of need

At the time of writing, `OpenFEPOPS` is the only publicly available implementation of the FEPOPS molecular similarity technique. Whilst used within industry and referenced extensively in literature, it has been unavailable to researchers as an open-source tool. We welcome contributions and collaborative efforts to enhance and expand OpenFEPOPS using the associated GitHub repository. It is therefore hoped that this will allow the technique to be used not only for

traditional small molecule molecular similarity, but also in new emerging fields such as protein design and featurisation of small- and macro-molecules for both predictive and generative tasks.

Brief software description

Whilst OpenFEPOPS has included functionality for descriptor caching and profiling of libraries, the core functionality of the package is descriptor generation and scoring.

Descriptor generation:

The OpenFEPOPS descriptor generation process as outlined in Figure 1 follows;

1. Tautomer enumeration
   - For a given small molecule, OpenFEPOPS uses RDKit (Landrum, 2013) to iterate over molecular tautomers. By default, there is no limit to the number of recoverable tautomers but a limit may be imposed which may be necessary if adapting the OpenFEPOPS code to large macromolecules and not just small molecules.
2. Conformer enumeration
   - For each tautomer, up to 1024 conformers are sampled by either complete enumeration of rotatable bond states (at the literature reported optimum increment of 90 degrees) if there are five or less rotatable bonds, or through random sampling of 1024 possible states if there are more than 5 rotatable bonds.
3. Defining feature points
   - The KMeans algorithm (Arthur & Vassilvitskii, 2007) is applied to each conformer of each tautomer to identify four (by default) representative or central points, into which the atomic information of neighbouring atoms is collapsed. As standard, the atomic properties of charge, logP, hydrogen bond donor, and hydrogen bond acceptor status are collapsed into four feature points per unique tautomer conformation. The RDKit package is used to calculate these properties with the iterative Gasteiger charges algorithm (Gasteiger & Marsili, 1980) applied to assign atomic charges, the Crippen method (Wildman & Crippen, 1999) used to assign atomic logP contributions, and hydrogen bond acceptors and donors identified with appropriate (Gillet et al., 1998) SMARTS substructure queries. These feature points are encoded to 22 numeric values (a FEPOP) comprising four points, each with four properties, and six pairwise distances between these points. With many FEPOPS descriptors collected from a single molecule through tautomer and conformer enumeration, this set of representative FEPOPS should capture every possible state of the original molecule.
4. Selection of diverse FEPOPS
   - From the collection of FEPOPS derived from every tautomer conformation of a molecule, the K-Medoid algorithm (Park & Jun, 2009) is applied to identify seven (by default) diverse FEPOPS which are thought to best capture a fuzzy representation of the molecule. These seven FEPOPS each comprise 22 descriptors each, totaling 154 32-bit floating point numbers or 616 bytes.

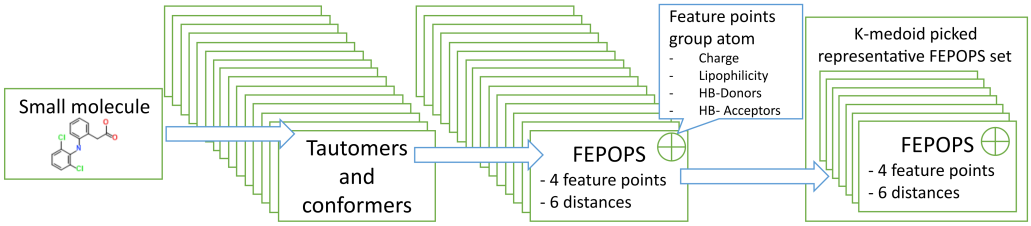

Figure 1: OpenFEPOPS descriptor generation showing the capture of tautomer and conformer information from a single input molecule.

Descriptor generation with OpenFEPOPS is a compute intensive task and as noted in literature, designed to be run in situations where large compound archives have had their descriptors pre-generated and are queried against relatively small numbers of new molecules for which descriptors are not known and are ad-hoc generated. To enable use in this manner, OpenFEPOPS provides functionality to cache descriptors through specification of database files, either in the SQLite or JSON formats.

Scoring and comparison of molecules based on their molecular descriptors

1. Sorting
   - With seven (by default) diverse FEPOPS representing a small molecule, the FEPOPS are sorted by ascending charge.
2. Scaling
   - Due to the different scales and distributions of features comprising FEPOPS descriptors, each FEPOP is centered and scaled according to observed mean and standard deviations of the same features within a larger pool of molecules. By default, these means and standard deviations have been derived from the DUDE (Mysinger et al., 2012) diversity set which captures known actives and decoys for a diverse set of therapeutic targets (See the Jupyter notebook 'Explore_DUDE_diversity_set.ipynb' in the source repository for further methods).
3. Scoring
   - The Pearson correlation coefficient is calculated for the scaled descriptors of the first molecule to the scaled descriptors of the second.

Literature highlights that the choice of the Pearson correlation coefficient leads to high background scores as it is highly unlikely to see little correlation between any molecule due to fundamental limitations of chemistry and geometry. Therefore, unrelated molecules are likely to have FEPOPS similarity scores higher than those encountered with more traditional techniques such as bitstring fingerprints and Tanimoto or Dice similarity measures.

The predictive performance of OpenFEPOPS was evaluated using the DUDE (Mysinger et al., 2012) diversity set. This dataset comprises eight protein targets accompanied by decoy ligands and known active ligands. For each target, actives were used as queries to retrieve all other actives. Retrieval rankings were assessed using the AUROC (Area Under Receiver Operating Characteristic) metric (Fawcett, 2006) and scores for each active averaged within targets to assign a final average AUROC score for each target. Table 1 shows the average AUROC scores for DUDE diversity set targets along with scores obtained using the popular Morgan 2, MACCS, and RDKit fingerprints as implemented in RDKit and scored using the Tanimoto distance metric. See the Jupyter notebook 'Explore_DUDE_diversity_set.ipynb' in the source repository for further methods and data availability using the FigShare service. All evaluated similarity techniques perform comparably with average AUROC scores of 0.723, 0.692, 0.687, and 0.701 for Morgan 2, MACCS, RDKit and OpenFEPOPS respectively. OpenFEPOPS achieves comparable performance to other metrics using 3D representations of molecules across a range of tautomer states which is in stark contrast to the approaches taken by the other connectivity and fingerprint-based methods. Diversity in similarity techniques allows potentially

interesting actives undiscoverable with one technique to be flagged and ranked highly by another, offering new routes to novelty, new chemistries, and efficacious leads from early-stage drug discovery efforts.

Target	Morgan 2	MACCS	RDKit	OpenFEPOPS
akt1	0.836	0.741	0.833	0.831
ampc	0.784	0.673	0.660	0.639
cp3a4	0.603	0.582	0.613	0.647
cxcr4	0.697	0.854	0.592	0.899
gcr	0.670	0.666	0.708	0.616
hivpr	0.780	0.681	0.759	0.678
hivrt	0.651	0.670	0.660	0.582
kif11	0.763	0.668	0.672	0.713

Table 1: Averaged AUROC scores by target and molecular similarity technique for the DUDE diversity set. Across all datasets, 19 small molecules out of 112,796 were excluded from analysis mainly due to issues in parsing to valid structures using RDKit.

Availability, usage and documentation

OpenFEPOPS has been uploaded to the Python Packaging Index under the name 'fepops' and as such is installable using the pip package manager and the command `pip install fepops`. With the package installed, entrypoints are used to expose commonly used OpenFEPOPS tasks such as descriptor generation and calculation on molecular similarity, enabling simple command line access without the need to explicitly invoke a Python interpreter. Whilst OpenFEPOPS may be used solely via the command line interface, a robust API is available and may be used within other programs or integrated into existing pipelines to enable more complex workflows. Extensive API documentation is available at https://justinykc.github.io/FEPOPS, along with a concise user-guide at https://justinykc.github.io/FEPOPS/readme.html

References

Arthur, D., & Vassilvitskii, S. (2007). K-means++ the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035. https://dl.acm.org/doi/abs/10.5555/1283383.1283494

Baber, J. C., Shirley, W. A., Gao, Y., & Feher, M. (2006). The use of consensus scoring in ligand-based virtual screening. *Journal of Chemical Information and Modeling*, *46*(1), 277–288. https://doi.org/10.1021/ci050296y

Cortés-Ciriano, I., Škuta, C., Bender, A., & Svozil, D. (2020). QSAR-derived affinity fingerprints (part 2): Modeling performance for potency prediction. *Journal of Cheminformatics*, *12*(1), 41. https://doi.org/10.1186/s13321-020-00444-5

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gasteiger, J., & Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, *36*(22), 3219–3228. https://doi.org/10.1016/0040-4020(80)80168-2

Gillet, V. J., Willett, P., & Bradshaw, J. (1998). Identification of biological activity profiles using substructural analysis and genetic algorithms. *Journal of Chemical Information and Computer Sciences*, *38*(2), 165–179. https://doi.org/10.1021/ci970431+

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x

Jenkins, J. L. (2013). Feature point pharmacophores (FEPOPS). *Scaffold Hopping in Medicinal Chemistry*, 155–174. https://doi.org/10.1002/9783527665143.ch10

Jenkins, J. L., Glick, M., & Davies, J. W. (2004). A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *Journal of Medicinal Chemistry*, *47*(25), 6144–6159. https://doi.org/10.1021/jm049654z

Landrum, G. (2013). *RDKit: Open-source cheminformatics*.

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, *55*(14), 6582–6594. https://doi.org/10.1021/jm300687e

Nettles, J. H., Jenkins, J. L., Williams, C., Clark, A. M., Bender, A., Deng, Z., Davies, J. W., & Glick, M. (2007). Flexible 3D pharmacophores as descriptors of dynamic biological space. *Journal of Molecular Graphics and Modelling*, *26*(3), 622–633. https://doi.org/10.1016/j.jmgm.2007.02.005

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Ripphausen, P., Nisius, B., & Bajorath, J. (2011). State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, *16*(9-10), 372–376. https://doi.org/10.1016/j.drudis.2011.02.011

Wildman, S. A., & Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, *39*(5), 868–873. https://doi.org/10.1021/ci990307l