


Improving reproducibility of cheminformatics workflows with chembl-downloader

Charles Tapley Hoyt ¹✉

¹ RWTH Aachen University, Institute of Inorganic Chemistry  ✉ Corresponding author

DOI: [10.21105/joss.08844](https://doi.org/10.21105/joss.08844)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Evan Spotte-Smith](#) 

Reviewers:

- [@PatWalters](#)
- [@dhimmel](#)

Submitted: 07 July 2025

Published: 10 September 2025

License

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

Statement of need

Many modern cheminformatics workflows derive datasets from ChEMBL ([Gaulton et al., 2017](#); [Zdrazil et al., 2023](#)), but few of these datasets are published with accompanying code for their generation. Consequently, their methodologies (e.g., selection, filtering, aggregation) are opaque, reproduction is difficult, and interpretation of results therefore lacks important context. Further, such static datasets quickly become out-of-date. For example, the current version of ChEMBL is v35 (as of December 2024), but ExCAPE-DB ([Sun et al., 2017](#)) uses v20, Deep Confidence ([Cortés-Ciriano & Bender, 2019](#)) uses v23, the consensus dataset from Isigkeit et al. (2022) uses v28, and Papyrus ([Béguignon et al., 2023](#)) uses v30. Therefore, there is a need for tools that provide reproducible bulk access to the latest (or a given) version of ChEMBL in order to enable researchers to make their derived datasets more transparent, updatable, and trustworthy.

State of the field

ChEMBL is typically accessed through its [application programming interface \(API\)](#), through its Python client ([Davies et al., 2015](#)), through its RDF platform ([Jupp et al., 2014](#)), or in bulk through its [file transfer protocol \(FTP\) server](#). However, APIs and their respective wrapper libraries are generally not efficient for querying and processing data in bulk due to dependency on network connection, remote server uptime and load, rate limits, and the need to paginate over results. Alternatively, bulk access is currently cumbersome for most potential users due to the need to download, set up, and connect to databases locally. Finally, third-party software such as [Pipeline Pilot](#), KNIME ([Berthold et al., 2009](#)), Galaxy ([The Galaxy Community, 2024](#)), and others reviewed by Warr (2012), that provide access to ChEMBL are often inflexible or inextensible due to being proprietary, closed source, or lacking approachable documentation.

Summary

This article introduces chembl-downloader, a Python package for the reproducible acquisition, access, and manipulation of ChEMBL data through its FTP server.

At a low-level, it uses a combination of the [pystow](#) Python software package and custom logic for the reproducible acquisition and pre-processing (e.g., uncompressing) of either the latest or a given version of most resources in the ChEMBL FTP server. These include relational database dumps (e.g., in SQLite), molecule lists (e.g., in SDF, TSV), pre-computed molecular fingerprints (e.g., in binary), a monomer library (e.g., in XML), and UniProt target mappings (e.g., in TSV).

At a mid-level, it provides utilities for accessing these files through useful data structures and functions such as querying the SQLite database with combination of Python's [sqlite3](#) library

and pandas (McKinney, 2010), parsing SDF files with RDKit (Landrum, n.d.), parsing TSVs with pandas, loading fingerprints with chemfp (Dalke, 2019), and parsing the monomer library with Python's xml library. The low- and mid-level utilities are kept small and simple such that they can be arbitrarily extended by users.

At a high-level, it maintains a small number of task-specific utilities. It contains several pre-formatted SQL queries to retrieve the bioactivities associated with a given assay or target, to retrieve the compounds mentioned in a publication or patent, to retrieve the names of all compounds, etc.

Case studies

A first case study demonstrates how the high-level utilities in chembl-downloader can be used to reproduce the dataset generation from Cortés-Ciriano & Bender (2019), highlight some of the methodological controversies (e.g., using arithmetic mean instead of geometric mean for pIC_{50} values), show the immense variability introduced by new datapoints in later versions of ChEMBL, and highlight the number of additional compounds added to each of its 24 target-specific datasets. Landrum & Riniker (2024) further investigated the impact of such aggregation. See the corresponding [Jupyter notebook](#).

A second case study demonstrates the value of having a reproducible script for identifying missing identifier mappings between molecules in ChEMBL and ChEBI (Hastings et al., 2016) via lexical mappings produced by Gilda (Gyori et al., 2022), which identified 4,266 potential mappings for curation e.g., in a workflow such as Biomappings (Hoyt et al., 2023). See the corresponding [Jupyter notebook](#).

A final case study demonstrates the utility of chembl-downloader by making pull requests to the code repositories corresponding to three popular cheminformatics blogs ([the RDKit Blog](#), [Practical Cheminformatics](#), and [Is Life Worth Living?](#)) to make the code more reproducible (where the source data was not available) and ChEMBL-version-agnostic:

- [greglandrum/rdkit_blog \(#5\)](#) for generating a substructure library
- [PatWalters/sfi \(#11\)](#) for calculating compounds' solubility forecast index, originally proposed by Hill & Young (2010)
- [PatWalters/jcamd_model_comparison \(#1\)](#) for comparing classification models
- [iwatobipen/playground \(#4\)](#) for parsing and using ChEMBL's monomer library
- [iwatobipen/playground \(#5\)](#) for analyzing chemical space of molecules from a set of patents

Additional external use cases can be found by [searching GitHub](#). Finally, several scholarly articles, including from the ChEMBL group itself, have used chembl-downloader in their associated code (Domingo-Fernández et al., 2023; Gadiya et al., 2023; Gorostiola González et al., 2024; Nisonoff et al., 2023; Schoenmaker et al., 2025; Zdrasil et al., 2023; Zhang et al., 2024).

Availability and usage

chembl-downloader is available as a package on [PyPI](#) with the source code available at <https://github.com/cthoht/chembl-downloader> and documentation available at <https://chembl-downloader.readthedocs.io>. The repository also contains an interactive Jupyter notebook tutorial and notebooks for the case studies described above.

Acknowledgements

The author would like to thank Yojana Gadiya and Jennifer HY Lin for helpful discussions and the NFDI4Chem Consortium (<https://www.nfdi4chem.de>) for support.

References

- Béquignon, O. J. M., Bongers, B. J., Jespers, W., IJzerman, A. P., Water, B. van der, & Westen, G. J. P. van. (2023). Papyrus: A large-scale curated dataset aimed at bioactivity predictions. *Journal of Cheminformatics*, 15(1), 3. <https://doi.org/10.1186/s13321-022-00672-x>
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1), 26–31. <https://doi.org/10.1145/1656274.1656280>
- Cortés-Ciriano, I., & Bender, A. (2019). Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.*, 59(3), 1269–1281. <https://doi.org/10.1021/acs.jcim.8b00542>
- Dalke, A. (2019). The chemfp project. *J. Cheminform.*, 11(1), 76. <https://doi.org/10.1186/s13321-019-0398-8>
- Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1), W612–W620. <https://doi.org/10.1093/nar/gkv352>
- Domingo-Fernández, D., Gadiya, Y., Mubeen, S., Healey, D., Norman, B. H., & Colluru, V. (2023). Exploring the known chemical space of the plant kingdom: Insights into taxonomic patterns, knowledge gaps, and bioactive regions. *Journal of Cheminformatics*, 15(1), 107. <https://doi.org/10.1186/s13321-023-00778-w>
- Gadiya, Y., Gribbon, P., Hofmann-Apitius, M., & Zaliani, A. (2023). Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3, 100069. <https://doi.org/10.1016/j.ailsci.2023.100069>
- Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrán-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.*, 45(D1), D945–D954. <https://doi.org/10.1093/nar/gkw1074>
- Gorostiola González, M., Béquignon, O. J. M., Manners, E., Gaulton, A., Mutowo, P., Dawson, E., Zdrazil, B., Leach, A. R., IJzerman, A. P., Heitman, L. H., & al., et. (2024). Excuse me, there is a mutant in my bioactivity soup! A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modelling. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv-2024-kxlgm>
- Gyori, B. M., Hoyt, C. T., & Steppi, A. (2022). Gilda: biomedical entity text normalization with machine-learned disambiguation as a service. *Bioinformatics Advances*. <https://doi.org/10.1093/bioadv/vbac034>
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, 44(D1), D1214–D1219. <https://doi.org/10.1093/nar/gkv1031>
- Hill, A. P., & Young, R. J. (2010). Getting physical in drug discovery: a contemporary

- perspective on solubility and hydrophobicity. *Drug Discov. Today*, 15(15), 648–655. <https://doi.org/10.1016/j.drudis.2010.05.016>
- Hoyt, C. T., Hoyt, A. L., & Gyori, B. M. (2023). Prediction and Curation of Missing Biomedical Identifier Mappings with Biomappings. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btad130>
- Isigkeit, L., Chaikuad, A., & Merk, D. (2022). A Consensus Compound/Bioactivity Dataset for Data-Driven Drug Design and Chemogenomics. *Molecules*, 27(8). <https://doi.org/10.3390/molecules27082513>
- Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., & Jenkinson, A. M. (2014). The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9), 1338–1339. <https://doi.org/10.1093/bioinformatics/btt765>
- Landrum, G. A. (n.d.). *RDKit: Open-source cheminformatics*. <http://www.rdkit.org>
- Landrum, G. A., & Riniker, S. (2024). Combining IC50 or ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5), 1560–1567. <https://doi.org/10.1021/acs.jcim.4c00049>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Nisonoff, H., Wang, Y., & Listgarten, J. (2023). Coherent blending of biophysics-based knowledge with bayesian neural networks for robust protein property prediction. *ACS Synthetic Biology*, 12(11), 3242–3251. <https://doi.org/10.1021/acssynbio.3c00217>
- Schoenmaker, L., Sastrokarijo, E. G., Heitman, L. H., Beltman, J. B., Jespers, W., & Westen, G. J. P. van. (2025). Towards assay-aware bioactivity model(er)s: Getting a grip on biological context. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv-2025-vnd2c>
- Sun, J., Jeliaskova, N., Chupakin, V., Golib-Dzib, J. F., Engkvist, O., Carlsson, L., Wegner, J., Ceulemans, H., Georgiev, I., Jeliaskov, V., Kochev, N., Ashby, T. J., & Chen, H. (2017). ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminform.*, 9(1), 1–9. <https://doi.org/10.1186/s13321-017-0203-5>
- The Galaxy Community. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res.*, 52(W1), W83–W94. <https://doi.org/10.1093/nar/gkae410>
- Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*, 26(7), 801–804. <https://doi.org/10.1007/s10822-012-9577-7>
- Zdrzil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., Veij, M. de, Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2023). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1), D1180–D1192. <https://doi.org/10.1093/nar/gkad1004>
- Zhang, H., Wu, J., Liu, S., & Han, S. (2024). A pre-trained multi-representation fusion network for molecular property prediction. *Information Fusion*, 103, 102092. <https://doi.org/10.1016/j.inffus.2023.102092>