MolSanitizer: A Python Pipeline for Preparing Large Databases of Small Molecules for Drug Discovery

Phong Lam[1], Szymon Pach[1], Ruth Brenk[2], Philip Ullmann[1], Flavio Ballante[3], Jens Carlsson[1], Israel Cabeza de Vaca[1]

[1]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.
[2]Department of Biomedicine, University of Bergen, Bergen, Norway.
[3]Chemical Biology Consortium Sweden, Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.

Aims: Accurate preparation of chemical databases is essential for structure-based drug discovery, yet many available tools rely on proprietary licensing, while open-source alternatives often underperform in key tasks or require extensive manual intervention. Here, we introduce a rule-based, Python pipeline for preparing small-molecule databases. MolSanitizer offers versatile functionalities, including protonation, tautomerization, unwanted substructure filtering, and conformational sampling. It supports popular file formats (SMILES, SDF, Mol2, PDBQT, DB2), enabling its integration into diverse structure-based drug discovery workflows.

Methods: MolSanitizer is built on RDKit and comprises three core modules. The first module predicts tautomeric and protonation states across different pH values using rule-based SMARTS reactions. The second module performs filtering through SMARTS matching to remove undesirable substructures. The third module focuses on conformational sampling, generating initial conformers with RDKit and exploring conformational space stochastically using the Torsion Library v3 [1].

Results: Preliminary tests on the DrugBank dataset showed good agreement with experimentally determined data. Conformational sampling benchmarks using the Platinum Diverse Set [2] reproduced 99.5% of bioactive conformations (≤2 Å RMSD). Enrichment analysis with DOCK3.8 on the DUDE-Z dataset [3] showed superior performance for MolSanitizer-generated conformations, achieving a mean logAUC of 18.7 compared to 15.1 for the original ZINC22 pipeline. Additionally, MolSanitizer is computationally efficient, running 21 times faster than RDKit ETKDGv3 [4] and matching the speed of the original conformer generator pipeline used in DOCK3.8.

Conclusions: MolSanitizer is an accurate, user-friendly, and open-source Python pipeline for chemical database preparation, offering high performance and flexibility as a viable alternative to proprietary workflows. MolSanitizer can be accessed through a web server at: https://carlssonlabtools.icm.uu.se/molsani.

References:

[1] Penner, P., Guba, W., Schmidt, R., Meyder, A., Stahl, M., Rarey, M., 2022. The Torsion Library: Semiautomated Improvement of Torsion Rules with SMARTScompare. J. Chem. Inf. Model. 62, 1644–1653.

[2] Friedrich, N.-O., Meyder, A., de Bruyn Kops, C., Sommer, K., Flachsenberg, F., Rarey, M., Kirchmair, J., 2017. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. J. Chem. Inf. Model. 57, 529–539.

[3] Stein, R.M., Yang, Y., Balius, T.E., O'Meara, M.J., Lyu, J., Young, J., Tang, K., Shoichet, B.K., Irwin, J.J., 2021. Property-Unmatched Decoys in Docking Benchmarks. J. Chem. Inf. Model. 61, 699–714.

[4] Wang, S., Witek, J., Landrum, G.A., Riniker, S., 2020. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. J. Chem. Inf. Model. 60, 2044–2058.