Assessment of the diversity of protein-bound ligand conformations and their representation with conformer ensembles

Nils-Ole Friedrich[1], Méliné Simsir[1,2], Kai Sommer[1], Matthias Rarey[1], Johannes Kirchmair[1]*

[1] *Universität Hamburg, Center for Bioinformatics, Hamburg, 20146, Germany*

[2] *Molécules thérapeutiques in silico (MTi), Université Paris Diderot, Sorbonne Paris Cité, Paris, France*

*kirchmair@zbh.uni-hamburg.de

Three-dimensional computational methods for guiding the discovery and optimization of bioactive small molecules rely on the accurate representation of the protein-bound conformations of ligands.[1] We have developed a cheminformatics pipeline for the fully automated identification and extraction of high-quality structures of protein-bound ligands from the PDB.[2,3] Importantly, among many other aspects, the support of the individual atom positions of ligands by the measured electron density is evaluated as part of this workflow. Using this software infrastructure, which we will present as part of this contribution, we have compiled a complete dataset of high-quality structures of protein-bound ligand conformations from the PDB, consisting of a total of 10,936 high-quality structures ("Sperrylite Dataset") of 4,548 unique ligands. This allowed us, for the first time, to conduct a comprehensive analysis of the diversity of protein-bound ligand conformations.

In total, we have studied the conformational variability of 91 drug-like molecules represented by a minimum of ten high-quality structures. We will show that a clear trend for the formation of few clusters of highly similar conformers is observed but that several interesting examples of small molecules that can adopt two or more distinct conformations when bound to different proteins exist, such as imatinib.

A diversified subset of this dataset was also used to assess how well leading free and commercial algorithms for conformer ensemble generation are able to represent bioactive conformations. We demonstrate that the differences in accuracy, computational cost and ensemble size are much smaller between commercial algorithms than those observed for free algorithms. RDKit generally achieved a favorable balance of accuracy, ensemble size and runtime among the seven tested free algorithms and its performance was comparable to that of mid-ranked commercial algorithms (median RMSD of 0.52 Å; measured between the bioactive conformation and the closest conformer in the ensemble). OMEGA obtained the best accuracy and speed among the eight tested commercial algorithms (median RMSD of 0.43 Å).

[1] Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.,* **2012**, *52*, 1146−1158

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

[3] Friedrich, N.-O., de Bruyn Kops, C., Flachsenberg, F., Sommer, K., Rarey, M., and Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.,* **2017**, *57*, 2719–2728.