Assessment of protein-ligand complexes structural diversity

Shinada, N.K., [2,3,4] Schmidtke, P., [4] de Brevern, A.G· [1,2]
[1]alexandre.debrevern@univ-paris-diderot.fr
[2] *INSERM, UMR_S 1134, DSIMB, Univ Paris, INTS, Laboratoire d'Excellence GR-Ex, Paris, France.*
[3] *SBX Corp., Tōkyō-to, Shinagawa-ku, Tōkyō, Japan*
[4] *Discngine SAS, Paris France.*

Nowadays, the RCSB Protein Data Bank (PDB) contained over 150,000 protein structures, these are critical to understand the underlying mechanism of protein-ligand binding. Grasping such knowledge goes through large-scale analysis where the quantity of data subsequently impacts their conclusions. High propensity of redundancy in the protein-ligand conformations is a well-known issue of the PDB. Using sequence-structure information and structural alignment on 104,777 protein-ligand complexes from the PDB, we've been able to classify those structures into three groups: (i) unique occurrence in the PDB, (ii) with multiple identical conformations and (iii) with distinct binding modes. This approach highlights mobile and rigid residues involved in the binding mechanism and the various binding modes adopted by specific ligands. Furthermore, 84% of our initial dataset has been clustered resulting in a 2.47-fold decrease with the consideration of only one representative conformation. This non-redundant dataset ensues a robust support for future large-scale analysis and machine learning applications in the drug design field.