**Supplementary Information**

Contents:

* Supplementary Methods.
* Table S1.
* Figure S1.
* Figure S2.
* Figure S3.

**Supplementary Methods**

Collection of motifs from Protein Data Bank systems with Rosetta

***Protein Data Bank File Collection***

Systems for protein-ligand motifs were obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). The catalog of protein-ligand systems that were available by RCSB at the time were downloaded. The list of files were obtained on January 26, 2023 with 198,160 files. Files were initially collected as a text list of comma-separated PDB codes (entire list seen in pdb\_list\_all\_newline\_separated.txt). For easier processing, file lists were separated by PDB codes starting with 1, 2, 3, 4, 5, 6, and 7-9 (pdb\_list\_1.txt, pdb\_list\_2.txt, pdb\_list\_3.txt, pdb\_list\_4.txt, pdb\_list\_5.txt, pdb\_list\_6.txt, pdb\_list\_789.txt). RCSB’s batch\_download.sh script was used to pull PDB files into our workspace from the comma-separated list files.

batch\_download.sh example command: ./batch\_download.sh -f pdb\_list\_789.txt -p

***Protein Data Bank File Preparation and Motif Collection***

The python script motif\_database\_generator\_updated\_1\_27\_23.py was run on a directory of protein-ligand PDB files to prepare the files and collect motifs. PDB files were modified to create copies with only one ligand per file.

motif\_database\_generator\_updated\_1\_27\_23.py uses no command line arguments, and works on all PDB files in the directory where the script is run in.

Argument files for collecting motifs with Rosetta were generated per single-ligand system. The below example is for running the Rosetta identify\_ligand\_motifs app on PDB 7BYU on ligand 1PG.

|  |
| --- |
| #allow ligands in system to exist and not have to worry about having params files of them included  -ignore\_unrecognized\_res  #Inputted PDB file  -s 7BYU\_1PG.pdb  #minimum cutoff for hydrogen bond score  -hb\_score\_cutoff -0.3  #minimum cutoff for water score  -water\_score\_cutoff -0.3  #minimum cutoff for packing score  -pack\_score\_cutoff -0.5  #optional flag to not print out generated motifs as pdb files  #Default value is true, and visualization can be good for debugging  #For large scale runs, it is worth including this flag and keeping it set as false  -output\_motifs\_as\_pdb false |

identify\_ligand\_motifs example command: /main/source/bin/identify\_ligand\_motifs.linuxgccrelease @7BYU\_1PG\_flags

***Protein-Ligand Duplicate Motif Removal***