**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 a modified PDB file of 7BYU on ligand 1PG, 7BYU\_1PG.pdb. The 7BYU\_1PG.pdb file had lines removed for ligand EDO.

|  |
| --- |
| #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: /rosetta/source/bin/identify\_ligand\_motifs.linuxgccrelease @7BYU\_1PG\_flags

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

Motifs from the PDB systems were compiled into a single motifs file. A pipeline was run on this file to remove motifs that are considered to be duplicates of other motifs in the list. Each motif in the list was compared against each other motif in the list to determine duplicate motifs. Motifs were considered duplicates if there was a complete match of atoms on the residue and ligand side of the motif, a root mean square distance (RMSD) within 0.8Å, and jump angle difference within 0.3 radians.

A Slurm script, remove\_duplicate\_motifs\_helper.job, calls the python scripts shuffle\_motif\_lines.py and remove\_duplicate\_motifs\_helper.py to iteratively remove duplicate motifs from the starting motifs list. The starting motifs list is broken into smaller segments of up to 100,000 motifs that are filtered for duplicate motifs, and then spliced back together. The spliced motifs file has its order of motifs randomized, and is then spliced for further duplicate motif removal. This process is repeated at least 5 times until duplicate motifs are no longer removed from the randomized smaller motifs lists. Variables at the start of remove\_duplicate\_motifs\_helper.job are defined to target the starting motifs file, path to Rosetta executable for duplicate motif removal, and the number of motifs per smaller motifs file.

|  |
| --- |
| #starting motifs file  file\_name=200000\_sample\_motifs\_file.motifs  #path to the executable for remove\_duplicate\_motifs  path\_to\_executable=/rosetta/source/bin/remove\_duplicate\_motifs.linuxgccrelease  #number of motifs per sub-file (ideally 100k-1M)  #there is a balance between too few motifs per sub to find duplicates and the jobs running too long (too many motifs per sub)  motifs\_per\_sub=100000 |

Below are example argument inputs for a file to run the Rosetta duplicate motif removal app, remove\_duplicate\_motifs.

|  |
| --- |
| #option to allow overwriting files when one currently exists  -out::overwrite true  #name of initial input motifs file that may contain duplicates  -motif\_filename 5000\_motifs.motifs  #Name of output motifs file with duplicates removed  -output\_file 5000\_motifs\_duplicates\_removed.motifs  #distance threshold for RMSD of atoms in angstroms  -duplicate\_dist\_cutoff 0.8  #angle threshold for atoms in radians  -duplicate\_angle\_cutoff 0.3 |

remove\_duplicate\_motifs example command:

/rosetta/source/bin/ remove\_duplicate\_motifs .linuxgccrelease @flags

***Enamine Library Preparation***

***Conformer Library Generation***

***Conformer Library Shape Database Generation***

***Conformer Shape Selection***