model_systems

Step	Function	Output
0	PDB ID from BindingDB Validation Set	PDB IDs from ValDS
1	Maps PDB IDs to Uniprot ACCs	ACCs
2	Download Uniprot xml file for each ACC	*xml files*
3	extract all PDB IDs related tob each xml file (Uniprot ACC)	many PDB IDs for each ACC
4	find PDB file in Hal cluster and create symlinks to working directory	*PDB file symlinks*
5	E.coli expression info from PDB file	E.coli exp. True/False for each PDB
6	extracts sequence from PDB file (SEQRES line in PDB file)	sequence for each PDB
7	extracts non-biopolymer chemical component (HET component) info from PDB file	ligand, cofactor, metal present in each PDB (3 letter code and full name)
8	extract method from PDB file (EXPDTA label)	structure determination method of PDB file
9	extracts SMILES or InChI for HET from Ligand Expo	SMILES and InChI for each HET in each PDB
10	gets ligand info from ChEMBL (uses BioServices and Pandas)	pkl files of following DataFrames: ● Bioactivity records for each Uniprot ACC ● Summary DataFrame for Ki, Kd, Kd1, Kd2, IC50 data ● DataFrame of approved drugs for each ACC
11	chembl data analysis	counts the elements in the DataFrames of each target protein(Uniprot ACC). Output is plk files for dataFrames: ● Number of Bioactivity Records vs Uniprot ACC ● Number of Bioactivity records with unique ligands(ingerdient compounds) vs Uniprot ACC ● Number of Ki/IC50/Kd type bioactivity records vs Uniprot type ● number of approved drugs vs Uniprot ACC