Binding Affinity Distribution in Protein-Ligand Pairs Between Substrates and Products

Kaitlyn Terrell – knt2039

Abstract

This paper delves into the intricate world of protein-ligand interactions, leveraging computational methodologies to uncover fundamental insights. Through the integration of data from the Protein Data Bank (PDB), UniProt, and Kyoto Encyclopedia of Genes and Genomes (KEGG), a comprehensive framework was developed for analyzing binding affinities between proteins and their associated ligands. Employing metabolomics approaches and molecular docking simulations with AutoDock Vina, we explored the binding preferences of substrates and products, shedding light on their relative affinities. Our findings reveal intriguing trends, suggesting that substrates exhibit higher binding affinity compared to products, thus enriching our understanding of protein function and potential therapeutic targets.

Introduction

There are currently 218,500 structures on the Protein Data Bank (PDB)[3]. 67,723 of these entries are experimentally determined Homo sapiens protein structures through crystallography methods such as X-ray diffraction. Many of these entries are variations or duplicates of the same protein. In the UniProt database, there exist 8,558 entries for Homo sapiens proteins with 3D structures. When mapped to the PDB dataset for experimentally determined protein structures, it has been found that the resulting mapped ID's can be refined to 7,605 proteins[1,2].

Alpha fold, the computational protein structure prediction software, considers the whole protein sequence from UniProt to generate structure predictions[2]. While confidence levels are typically high for predicting "important" domains of the protein, when predicting the whole structure, the confidence tends to be low[3]. Occasionally, we see that ligand binding pockets found through experimental crystal structures have different locations from those predicted by alpha fold, outside the domain of the crystal structure[7,8]. While there are missing parts in both the alpha fold predictions and the experimental crystal structures, targeting only fragments of the sequence / individual domains, we want to know how much these missing parts impact ligand binding. Therefore, we exclude computationally determined structures in our dataset.

The development of a targetome to determine the products and substrates associated with each protein reaction, and ultimately the binding affinities of each protein-ligand pair, will provide further insight into protein classification, drug discovery, and therapeutics[5]. We hypothesize that the resulting binding affinity predictions, produced by the AutoDock Vina[8] software via PyMol, in protein-ligand pairs will be higher for substrates than for products.

Methods

Dataset Preparation and Retrieval

A metabolomics approach is used to develop a targetome of protein-ligand binding pairs, and assess the binding affinity of these substrates and products[5,9]. Development of a mapped file to KEGG enzyme and metabolite entries allows us to see a bigger picture of how many reactions are documented between metabolites and proteins[4]. Utilizing this dataset to retrieve corresponding compound structures from PubChem, and protein structures from PDB, we teamed up with Gil Speyer at Arizona State University to determine binding affinity for each pair using AutoDock Vina.

Using the mapped file of 7,605 UniProt and PDB ID's for proteins with experimentally determined crystal structures, I was able to utilize the packages in python to retrieve corresponding EC numbers from enzyme entries in the Kyoto Encyclopedia of Genes and Genomes (KEGG).

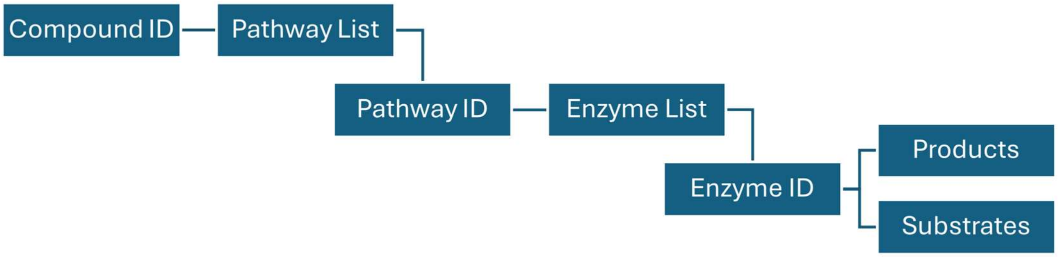


Figure 1. Flowchart for Glucose Control Pipeline

With Glucose as the control, the compound entry data was retrieved and parsed using the BioServices KEGG API in Python to yield a list of associated metabolic pathways in the form of Map ID's. The BS4 BeautifulSoup html parser was then used to retrieve EC numbers for enzymes/proteins associated with each pathway.

The reactions listed in reaction entries for each pathway were not reliable to determine the products and substrates, as all KEGG reactions are in equilibrium. Therefore, the enzyme entries were used, where reactions listed within have defined substrates and products, listed as compound ID's. The resulting dataframe includes all pathways associated with glucose listed as Map ID's, each subsequent enzyme for each pathway listed as EC numbers, and each substrate and product for each enzyme entry, listed as compound ID's. Prior to determining the products and substrates for this pipeline, the dataset was refined to remove any NaN EC's that were not mapped successfully to the input file.

The pipeline for the control compound, Glucose (KEGG Compound ID: C00031) was repurposed into a global analysis for all proteins in the mapped input file. This global analysis pipeline took a different approach, where EC numbers were extracted directly from each UniProt entry using the Bioservices UniProt API, along with Pandas and StringIO to read the file and search for entries. Each EC number was mapped to the corresponding UniProt ID. However, there were 5 distinct formats that these could occur as, classified herein as: None, Full, Partial -, Partial n, and Multiple[10]. UniProt entries occasionally list orthologs to the recommended protein names, which are represented as EC numbers, and reported in the dataset as Alternative EC numbers.

Type	**Example**	**Description**
"None"	None	indicating no EC number detected in the UniProt Entry page
"Full"	#.#.#.#	normal representation of EC number
"Partial"	#.#.#.-	unknown catalytic activity of the protein
"Partial"	#.#.#.n#	protein catalyzes known but uncatalogued rxn
"Multiple"	#.#.#.#, #.#.#.#	indicating multiple entries (full and/or partial)

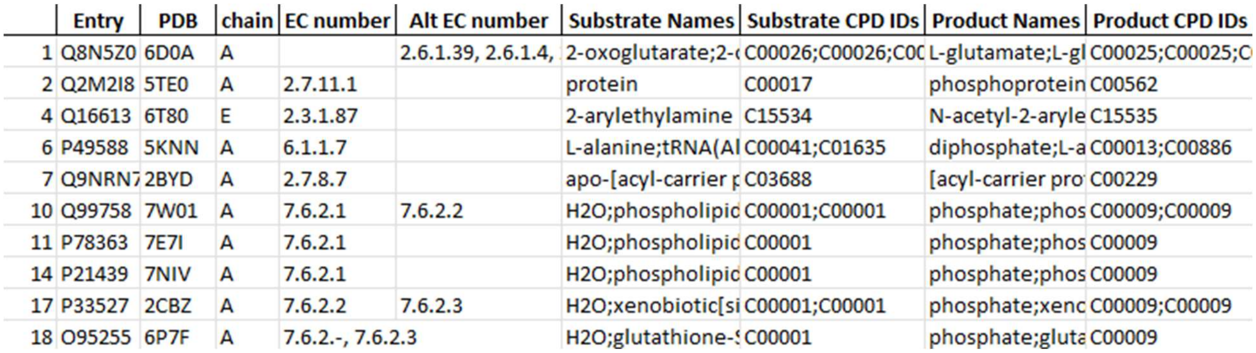
Figure 2. Classification table for EC number variations and descriptions.[10]

The addition of "type" to the dataframe helps for easier analysis (Figure 2). Statistical analysis was performed on differences of EC numbers reported, considering category, frequency, and alternative vs. recommended. Visualization of type distribution in the EC number column per 7606 protein was represented as percentages in a pie chart for visualization (Figure 7).

Data Analysis and Visualization

A plot comparing between EC number and Alt EC number types was produced, represented as a bar plot visualizing distribution of type for each type combo between EC number : Alt EC number columns. There are 16 type combinations, but there are 0 occurrences for 7/16 combinations: Full-Multiple, Multiple-Multiple, Multiple-Full, Multiple-None, Multiple-Partial, None-Multiple, and Partial-Multiple (Figure 8A).

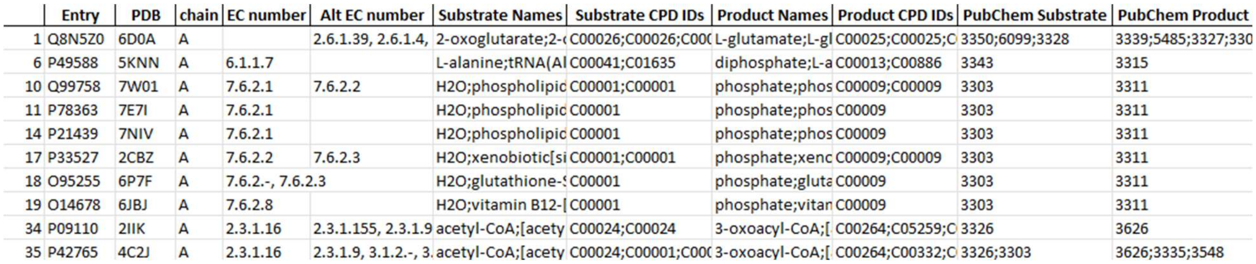
Further filtering of the dataset for entries with None in both columns (EC number, Alt EC number), as well as any combination of Partial and None entries was done. This provides a clearer foundation for determining substrates and products for each entry, as these could not be retrieved from EC numbers in the Partial or None categories. We can see in the dataframe produced by the pipeline below, the entries will not be a combination of Nan-Nan values, but if there is NaN entry in one column, the other will surely contain a Full entry (Figure 3).



	Entry	PDB	chain	EC number	Alt EC number	Substrate Names	Substrate CPD IDs	Product Names	Product CPD IDs
1	Q8N5Z0	6D0A	A		2.6.1.39, 2.6.1.4,	2-oxoglutarate;2-(	C00026;C00026;C00	L-glutamate;L-gl	C00025;C00025;C
2	Q2M2I8	5TE0	A	2.7.11.1		protein	C00017	phosphoprotein	C00562
4	Q16613	6T80	E	2.3.1.87		2-arylethylamine	C15534	N-acetyl-2-aryle	C15535
6	P49588	5KNN	A	6.1.1.7		L-alanine;tRNA(Al	C00041;C01635	diphosphate;L-a	C00013;C00886
7	Q9NRN7	2BYD	A	2.7.8.7		apo-[acyl-carrier p	C03688	[acyl-carrier pro	C00229
10	Q99758	7W01	A	7.6.2.1	7.6.2.2	H2O;phospholipid	C00001;C00001	phosphate;phos	C00009;C00009
11	P78363	7E7I	A	7.6.2.1		H2O;phospholipid	C00001	phosphate;phos	C00009
14	P21439	7NIV	A	7.6.2.1		H2O;phospholipid	C00001	phosphate;phos	C00009
17	P33527	2CBZ	A	7.6.2.2	7.6.2.3	H2O;xenobiotic[si	C00001;C00001	phosphate;xenc	C00009;C00009
18	O95255	6P7F	A	7.6.2.-, 7.6.2.3		H2O;glutathione-S	C00001	phosphate;gluta	C00009

Figure 3. A subset of the first 10 Protein entries for the Refined_MAP_2.xlsx output

Analysis of binding affinity is done by Gil Speyer at ASU using the molecular docking software, AutoDock Vina, to determine affinity of the binding pairs. Dataset preparation for this analysis required addition of PDB-CCD ID’s, retrieved from compound entries in KEGG using the Bioservices KEGG API to parse out identifiers from alternative databases. Similarly, PubChem ID’s were retrieved from the same section in each KEGG compound entry using the same method and added to the dataset. The PubChem_MAP file was further refined to remove any NaN rows (no PubChem ID listed). Refined PubChem MAP file removes entries with NaN in both product and substrate PubChem ID columns. The output file is named PubChem_MAP_Refined_2.xlsx (Figure 4).

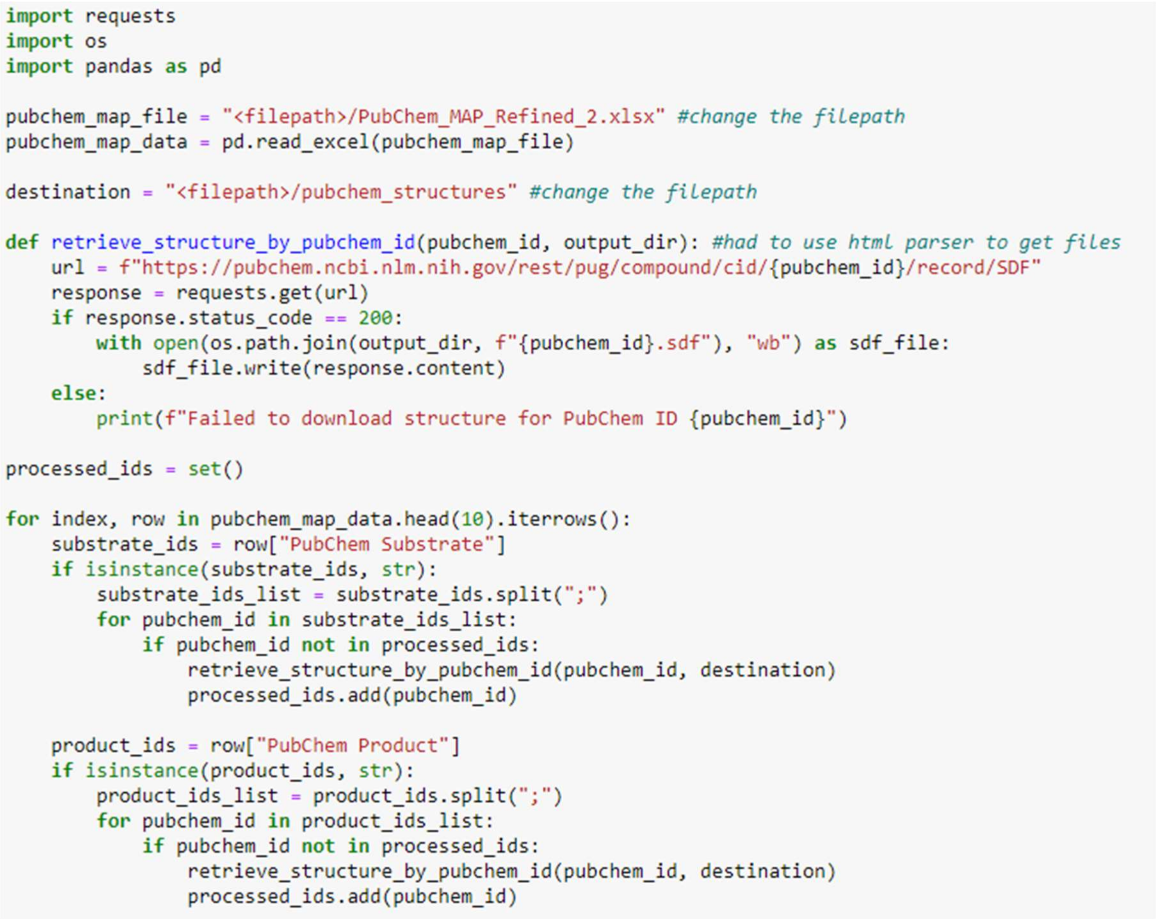


	Entry	PDB	chain	EC number	Alt EC number	Substrate Names	Substrate CPD IDs	Product Names	Product CPD IDs	PubChem Substrate	PubChem Product
1	Q8N5Z0	6D0A	A		2.6.1.39, 2.6.1.4,	2-oxoglutarate;2-(	C00026;C00026;C00	L-glutamate;L-gl	C00025;C00025;C	3350;6099;3328	3339;5485;3327;330
6	P49588	5KNN	A	6.1.1.7		L-alanine;tRNA(Al	C00041;C01635	diphosphate;L-a	C00013;C00886	3343	3315
10	Q99758	7W01	A	7.6.2.1	7.6.2.2	H2O;phospholipid	C00001;C00001	phosphate;phos	C00009;C00009	3303	3311
11	P78363	7E7I	A	7.6.2.1		H2O;phospholipid	C00001	phosphate;phos	C00009	3303	3311
14	P21439	7NIV	A	7.6.2.1		H2O;phospholipid	C00001	phosphate;phos	C00009	3303	3311
17	P33527	2CBZ	A	7.6.2.2	7.6.2.3	H2O;xenobiotic[si	C00001;C00001	phosphate;xenc	C00009;C00009	3303	3311
18	O95255	6P7F	A	7.6.2.-, 7.6.2.3		H2O;glutathione-S	C00001	phosphate;gluta	C00009	3303	3311
19	O14678	6JBJ	A	7.6.2.8		H2O;vitamin B12-[	C00001	phosphate;vitan	C00009	3303	3311
34	P09110	2IIK	A	2.3.1.16	2.3.1.155, 2.3.1.9	acetyl-CoA;[acety	C00024;C00024	3-oxoacyl-CoA;[	C00264;C05259;C	3326	3626
35	P42765	4C2J	A	2.3.1.16	2.3.1.9, 3.1.2.-, 3.	acetyl-CoA;[acety	C00024;C00001;C00	3-oxoacyl-CoA;[	C00264;C00332;C	3326;3303	3626;3335;3548

Figure 4. Subset of first 10 Protein entries for the PubChem_MAP_Refined_2.xlsx output

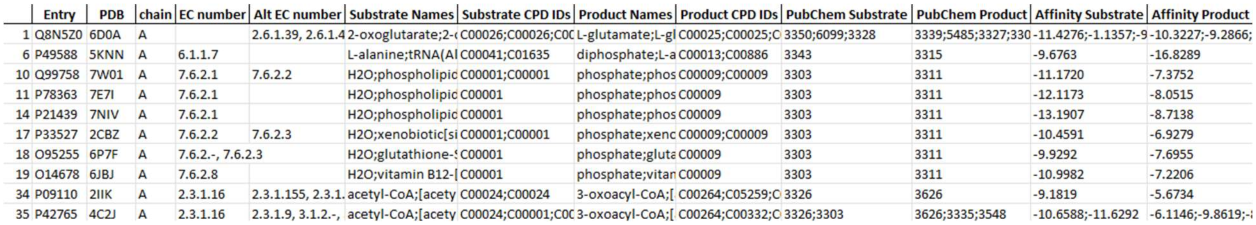
The following code was provided to Gil Speyer at ASU along with the output .xlsx file from the above chunk, listing all PubChem IDs for target substrates. The output PubChem_MAP.xlsx file above had 240 individual PubChem IDs, for which automation of PubChem structure file retrieval and download was performed on a subset of 10 metabolites to ensure functionality, and handed off to Gil Speyer for use in the binding affinity analysis. Automation of PDB structures from all entries in the original dataset was also developed, although these files have previously been obtained by Speyer (Figure 5).

Speyer then used these structure files along with the PDB structure files previously obtained to make the AutoDock Vina binding affinity predictions for each protein-ligand pair. He provided me with an output file called table_data.txt. From this file, I was able to parse the PDB ID from the chain, and the PubChem ID to map the binding affinities to each associated row entry. A data frame was used to access the data from this step for computational ease. Then, PubChem_MAP_Refined_2.xlsx was used as the input top map the data frame of binding affinities to new “Affinity Product” and Affinity Substrate” columns. This was further refined to remove any NaN values that occur in both new columns, as well as non-numerical formatting errors that were mapped as a result of the AutoDock Vina prediction. The resulting file is named Affinity_MAP_Refined.xlsx (Figure 6).



```python
import requests
import os
import pandas as pd

pubchem_map_file = "<filepath>/PubChem_MAP_Refined_2.xlsx" #change the filepath
pubchem_map_data = pd.read_excel(pubchem_map_file)

destination = "<filepath>/pubchem_structures" #change the filepath

def retrieve_structure_by_pubchem_id(pubchem_id, output_dir): #had to use html parser to get files
    url = f"https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/{pubchem_id}/record/SDF"
    response = requests.get(url)
    if response.status_code == 200:
        with open(os.path.join(output_dir, f"{pubchem_id}.sdf"), "wb") as sdf_file:
            sdf_file.write(response.content)
    else:
        print(f"Failed to download structure for PubChem ID {pubchem_id}")

processed_ids = set()

for index, row in pubchem_map_data.head(10).iterrows():
    substrate_ids = row["PubChem Substrate"]
    if isinstance(substrate_ids, str):
        substrate_ids_list = substrate_ids.split(";")
        for pubchem_id in substrate_ids_list:
            if pubchem_id not in processed_ids:
                retrieve_structure_by_pubchem_id(pubchem_id, destination)
                processed_ids.add(pubchem_id)

    product_ids = row["PubChem Product"]
    if isinstance(product_ids, str):
        product_ids_list = product_ids.split(";")
        for pubchem_id in product_ids_list:
            if pubchem_id not in processed_ids:
                retrieve_structure_by_pubchem_id(pubchem_id, destination)
                processed_ids.add(pubchem_id)
```

Figure 5. PubChem structure file scraper to retrieve and download structure files.



	Entry	PDB	chain	EC number	Alt EC number	Substrate Names	Substrate CPD IDs	Product Names	Product CPD IDs	PubChem Substrate	PubChem Product	Affinity Substrate	Affinity Product
1	Q8N5Z0	6D0A	A		2.6.1.39, 2.6.1.4	2-oxoglutarate;2-(	C00026;C00026;C0(	L-glutamate;L-gl	C00025;C00025;C	3350;6099;3328	3339;5485;3327;330	-11.4276;-1.1357;-9	-10.3227;-9.2866;
6	P49588	5KNN	A	6.1.1.7		L-alanine;tRNA(Al	C00041;C01635	diphosphate;L-a	C00013;C00886	3343	3315	-9.6763	-16.8289
10	Q99758	7W01	A	7.6.2.1	7.6.2.2	H2O;phospholipic	C00001;C00001	phosphate;phos	C00009;C00009	3303	3311	-11.1720	-7.3752
11	P78363	7E7I	A	7.6.2.1		H2O;phospholipic	C00001	phosphate;phos	C00009	3303	3311	-12.1173	-8.0515
14	P21439	7NIV	A	7.6.2.1		H2O;phospholipic	C00001	phosphate;phos	C00009	3303	3311	-13.1907	-8.7138
17	P33527	2CBZ	A	7.6.2.2	7.6.2.3	H2O;xenobiotic[si	C00001;C00001	phosphate;xenc	C00009;C00009	3303	3311	-10.4591	-6.9279
18	O95255	6P7F	A	7.6.2.-, 7.6.2.3		H2O;glutathione-!	C00001	phosphate;gluta	C00009	3303	3311	-9.9292	-7.6955
19	O14678	6JBJ	A	7.6.2.8		H2O;vitamin B12-[	C00001	phosphate;vitan	C00009	3303	3311	-10.9982	-7.2206
34	P09110	2IIK	A	2.3.1.16	2.3.1.155, 2.3.1.	acetyl-CoA;[acety	C00024;C00024	3-oxoacyl-CoA;[	C00264;C05259;C	3326	3626	-9.1819	-5.6734
35	P42765	4C2J	A	2.3.1.16	2.3.1.9, 3.1.2.-,	acetyl-CoA;[acety	C00024;C00001;C0(	3-oxoacyl-CoA;[	C00264;C00332;C	3326;3303	3626;3335;3548	-10.6588;-11.6292	-6.1146;-9.8619;-

Figure 6. Affinity_MAP_Refined.xlsx file including the refined Affinity scores for protein-ligand pairs.

The more negative the affinity score indicates a higher binding strength with the protein of interest. Therefore, we aim to look at the distribution of substrate vs product affinity with each protein to determine if there is variation between the ligand type. We hypothesize that substrate-protein binding strength will be higher (more negative) than product-protein binding affinity. This mapping file was then used to perform analysis of affinity distribution between products and substrates for each protein.

Results and Discussion

EC number Type Distribution

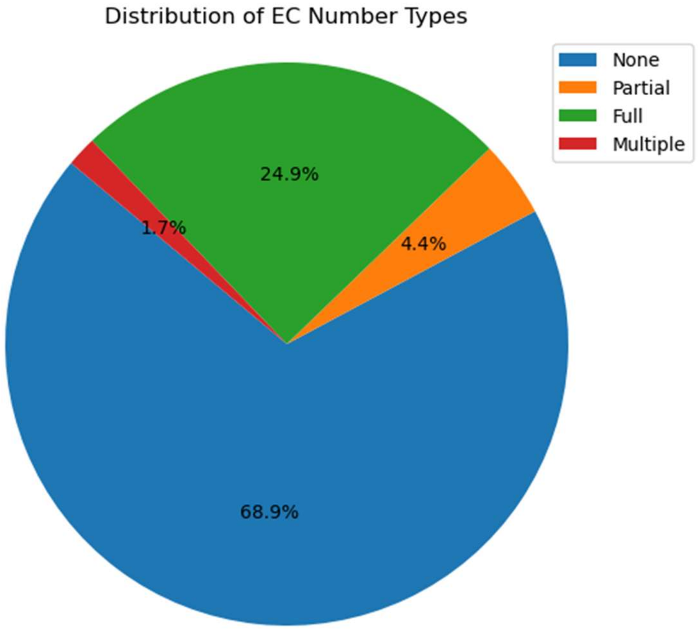
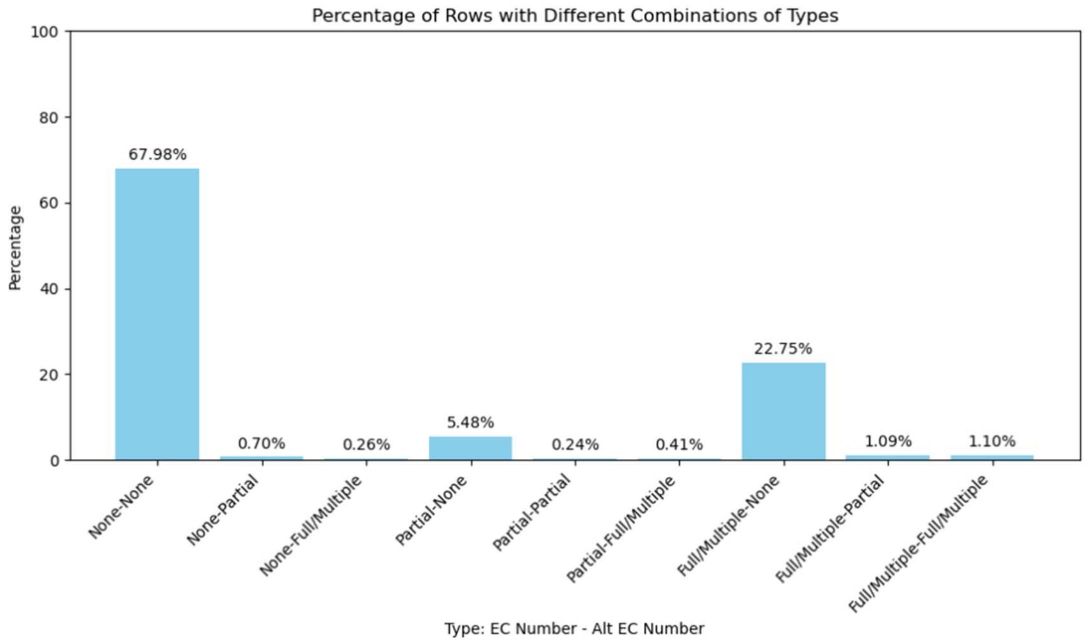
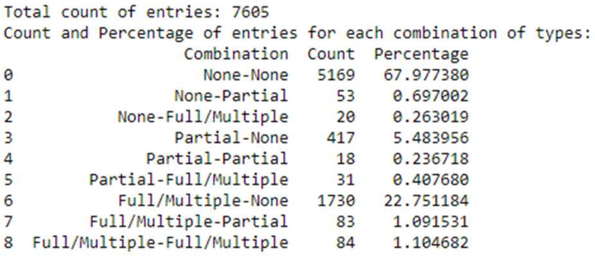


Figure 7. Pie chart for None, Partial and Full+Multiple EC number entries in Mapped file of 7606 proteins, just for the EC number column.

In the pie chart for EC number type distribution, we can see here that majority, 68.9%, of the entries are of type “None”, but almost 25% are “Full” entries that can be used for further analysis of ligand binding (Figure 7). We know that “partial” type represents either unknown catalytic activity of the protein, or that the protein catalyzes known but uncatalogued reactions (Figure2)[10]. Because of this, the “partial” type EC numbers cannot be accurately navigated to, since many encompass a larger family of enzymes. Therefore, “partial” EC numbers are not considered in further analysis. However, while “multiple” entries occur less often, many are still usable, and were included in the analysis, omitting any partial EC numbers in the list. We see that for proteins with multiple EC numbers, there is typically a high similarity between function, reactions, and associated products and substrates listed for each.





```
Total count of entries: 7605
Count and Percentage of entries for each combination of types:
                  Combination  Count  Percentage
0                   None-None   5169   67.977380
1                None-Partial     53    0.697002
2          None-Full/Multiple     20    0.263019
3                Partial-None    417    5.483956
4             Partial-Partial     18    0.236718
5       Partial-Full/Multiple     31    0.407680
6          Full/Multiple-None   1730   22.751184
7       Full/Multiple-Partial     83    1.091531
8  Full/Multiple-Full/Multiple     84    1.104682
```

Figure 8A. Bar plot showing type pairs for EC number percentages per protein. **B.** Counts and percentages of type pairs for EC number percentages per protein.

In the bar chart, we can see that majority of the entries have no EC number, and therefore will have no resulting substrate/product combination. The highest categories for entries with existing EC numbers was Full/Multiple-None, and Partial-None, which suggests no orthologs have EC numbers for most entries. The type combos we are interested in are any that are not None-None, but primarily the full/multiple or partial in either EC number column or Alt EC number column. Where, for our analysis, we can find a product/substrate combination listed for each protein of interest here (Figure 8A). We also can see the counts for each type pairing corresponding to the bar chart, where 7/16 combinations were removed due to zero percent occurrence (Figure 8B).

Binding Affinity Analysis

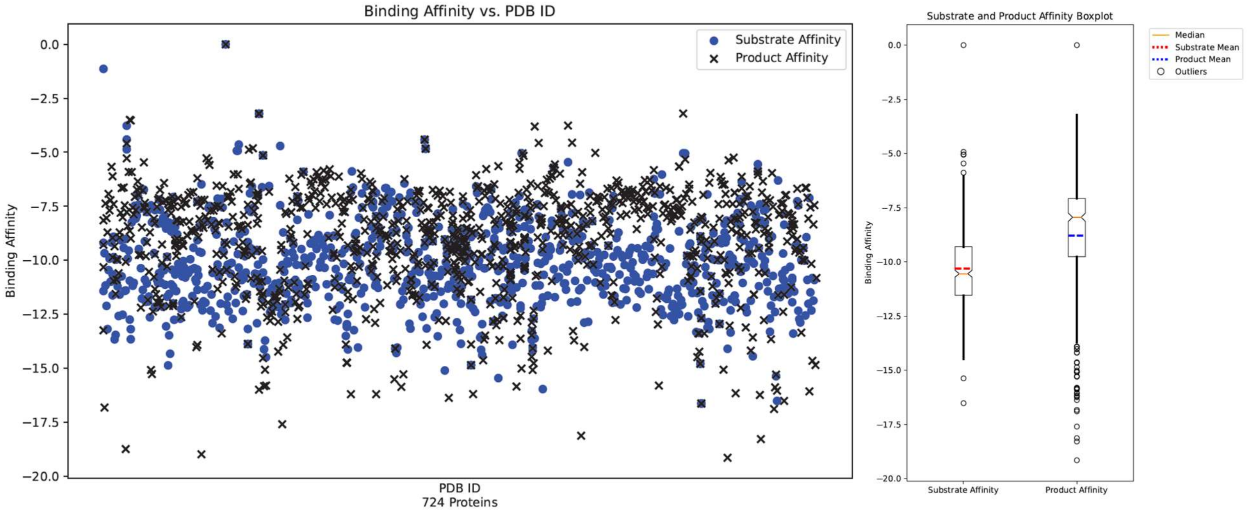


Figure 9. A. Scatter plot of binding affinity for substrate (O) and product (X) binding affinity for ligand-protein pairs. **B.** Box and whisker plots of substrate and product affinity with median, mean and outliers (o).

Multicolored affinity scores per PDB IDs for all 724 IDs is represented in the scatterplot above. The plot is represented with no PDB labels for clarity, Substrate Affinity (Blue O) and Product Affinity (Black X) scores represented in scatter plot. The median distribution of product affinity sits around -8.0, with median substrate affinity at -10.5 (Figure 9A). The box and whisker plot is scaled adjacent to the scatter plot to more easily visualize the distribution of substrate and product median affinity, showing the IQR (interquartile range) between the 25th and 75th percentiles, outliers are considered at 1.5xIQR above and below the whiskers. Median indicated by thin orange line and notch, means displayed by thick dashed line, outliers are marked as O (Figure 9B). This supports that the substrate affinity is higher (more negative) than product affinity between ligand-protein pairs. The product plot shows several outliers that are at higher affinity, but do not belong to the majority of product scores.

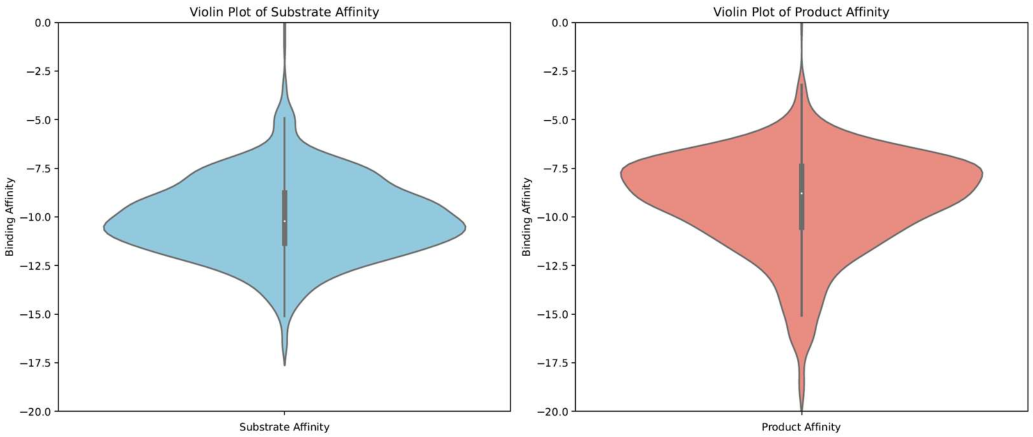


Figure 10. Violin plot to show distribution of substrate vs product affinity scores.

The violin plot of substrate vs product binding affinities for each protein-ligand pair for all 724 proteins. This distribution shows that the median affinity for substrates is higher than for products, as the more negative the score indicates a higher binding strength for that protein-ligand pair. This supports our hypothesis that the substrate binding affinity is higher than product binding affinity (Figure 10).

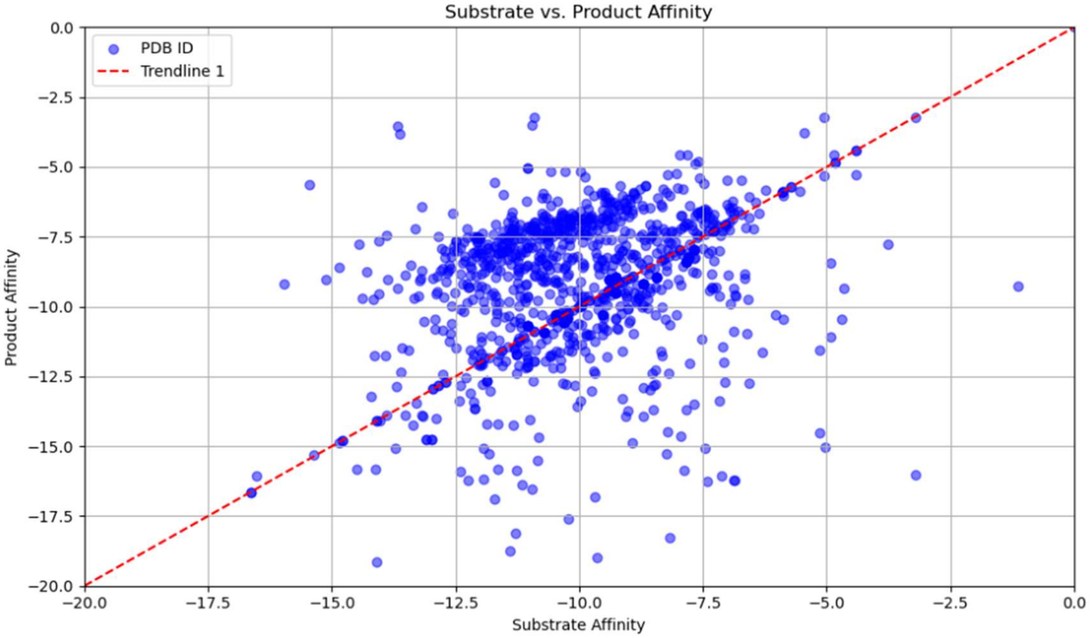


Figure 11. Correlation plot of substrate vs. product affinities per protein in 724 proteins.

The paired scatter plot of product vs substrate Affinities for 724 Protein-Ligand pairs from Affinity_MAP_Refined.xlsx shows 2 linear trends here. We can see here that the central trendline demonstrates the correlation between substrate and product affinity scores per protein for ligands that share similar affinity scores. While the higher trendline shows that the binding affinity between the substrate-protein pair is higher than the product-protein pair for the same protein. The outliers of products that have higher affinity, located below the central trendline, will be isolated and analyzed in future work to determine the molecular mechanisms associated with their proteins (Figure 11). If we compare this plot to the box and whisker plots for substrate and product affinities, we can see that the outliers for the high product affinities can be clearly observed below the central trendline (Figure 9B).

Conclusion

Preparation of the dataset required retrieval of enzymatic reactions based on EC numbers available from the KEGG database, production of a protein mapping file for products and substrates from these protein-ligand interactions, and retrieval of

PubChem structures for each ligand. From the final mapping dataset, we were able to make predictions of binding affinities for each protein-ligand pair using AutoDock Vina.

Our investigation demonstrates that substrate-protein interactions exhibit significantly higher binding affinities compared to product-protein interactions, supporting our initial hypothesis. This finding underscores the role of substrates in enzymatic reactions and metabolic processes, highlighting their importance in driving these protein-ligand interactions.

Furthermore, the observation of the outliers characterized by high binding affinities between product-protein pairs reveals a potential subset of proteins with functional or structural variation from the majority. Moving forward, we aim to perform a more in-depth analysis of the proteins with high product binding affinities. These outliers provide an intriguing motivation for further investigation into their significance and molecular function.

References

1. Ovanessians A, Snow C, Jennewein T, Sarkar S, Speyer G, Klein-Seetharaman J. Creation of a Curated Database of Experimentally Determined Human Protein Structures for the Identification of Its Targetome. Pac Symp Biocomput. 2024;29:291-305. PMID: 38160287.
2. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42. doi: 10.1093/nar/28.1.235. PMID: 10592235; PMCID: PMC102472.
4. Kanehisa M. KEGG Bioinformatics Resource for Plant Genomics and Metabolomics. Methods Mol Biol. 2016;1374:55-70. doi: 10.1007/978-1-4939-3167-5_3. PMID: 26519400.
5. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. J Am Soc Mass Spectrom. 2016 Dec;27(12):1897-1905. doi: 10.1007/s13361-016-1469-y. Epub 2016 Sep 13. PMID: 27624161; PMCID: PMC5110944.
6. Yang T, Hui R, Nouws J, Sauler M, Zeng T, Wu Q. Untargeted metabolomics analysis of esophageal squamous cell cancer progression. J Transl Med. 2022 Mar 14;20(1):127. doi: 10.1186/s12967-022-03311-z. PMID: 35287685; PMCID: PMC8919643.
7. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model. 2021 Aug 23;61(8):3891-3898. doi: 10.1021/acs.jcim.1c00203. Epub 2021 Jul 19. PMID: 34278794; PMCID: PMC10683950.
8. Seeliger D, de Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. J Comput Aided Mol Des. 2010 May;24(5):417-22. doi: 10.1007/s10822-010-9352-6. Epub 2010 Apr 17. PMID: 20401516; PMCID: PMC2881210.
9. Tian Y, Wan N, Zhang H, Shao C, Ding M, Bao Q, Hu H, Sun H, Liu C, Zhou K, Chen S, Wang G, Ye H, Hao H. Chemoproteomic mapping of the glycolytic targetome in cancer cells. Nat Chem Biol. 2023 Dec;19(12):1480-1491. doi: 10.1038/s41589-023-01355-w. Epub 2023 Jun 15. PMID: 37322158.
10. UniProt. "UniProt Release Notes: October 23, 2007." UniProt Knowledgebase. https://www.uniprot.org/release-notes/2007-10-23-release.