Analyzing Patterns of Computational Similarity between Protein Kinase Ligands

Jack Ringer

July 2025

1 Abstract

This work investigates whether there is a relationship between the 2D computational similarity of ligands and their activity within specific human protein kinase groups. Using data from ChEMBL binding assays, the distribution of pairwise Tanimoto similarity coefficients values computed between all protein kinase ligands was compared against the distribution of pairwise similarity values computed with respect to ligands active within a specific protein kinase group. Except for the CK1 group, no significant group-specific differences were found. These results suggest there is limited utility of 2D similarity metrics for identifying ligand selectivity across a majority of protein kinase groups. However, given the many confounders that exist when performing large scale computational analyses of ChEMBL bioassay data these results are not definitive, and limitations as well as potential follow-ups are discussed in detail. All code developed as part of this project can be found on GitHub: https://github.com/Jack-42/ligandActivityAnalysis.

2 Introduction and Background

This work explores the relationship between 2D ligand similarity and the activity of these ligands with respect to major protein kinase families, and in particular whether ligands which are active within a particular protein kinase group are more similar to one another than protein kinase ligands generally. The following sections provide a brief overview of protein kinases and their classification, molecular similarity, the project's relevance to drug discovery, and related works.

2.1 Protein Kinases

Given their (nearly impossible-to-overstate) importance in drug discovery, medicine, and biology + chemistry broadly, an immense amount of effort has been put into classifying proteins. One of the most significant protein families studied by drug discovery researchers is the protein kinase family. These protein kinases (PKs) are enzymes which modify the function of other proteins via phosphorylation [8]. PKs are involved in many important regulatory roles throughout the cell, and their dysregulation is linked to many types of cancer as well as immune, neurological and infectious diseases [5]. Although all PKs perform phosphorylation, they do not all perform the same function (e.g., some PKs will target different protein domains than others).

One of the first kinase classifications was published by Manning and other researchers in 2002 [17]. Their research has resulted in the establishment of 8 major groups within the human kinome, which include: AGC, CAMK, CK1, CMGC, STE, TK, TKL, Other, as well as 13 atypical families [9]. Generally speaking, these groups (as well as the families and subfamilies they contain) have been classified based on sequence similarity, evolutionary conservation, and known functions. The exceptions are the "Atypical" and "Other" groups, which serve to classify proteins that don't fit into the other major groups. A phylogenetic tree of the human kinome is shown in Figure 1.

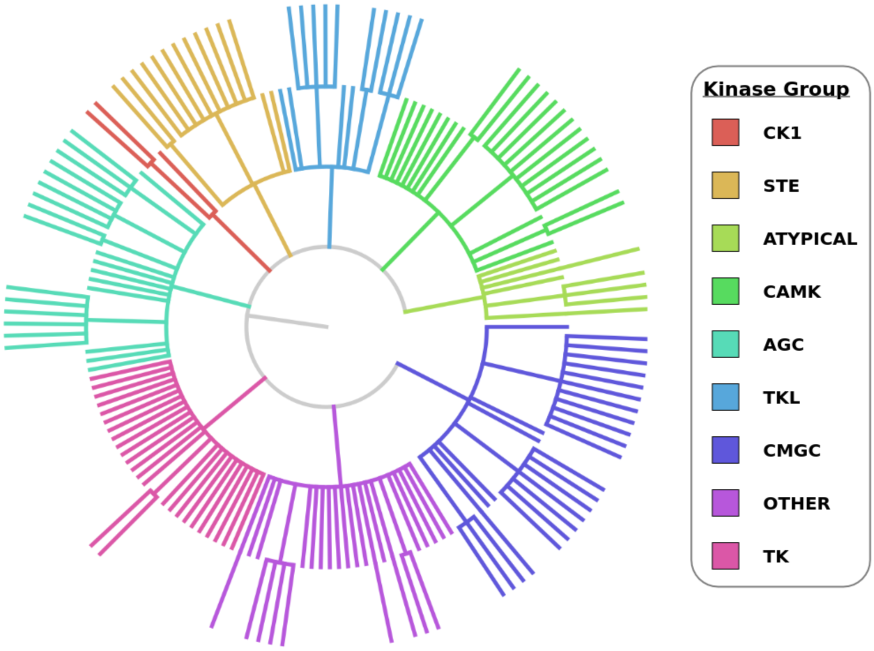

Figure 1: Phylogenetic Tree of the Human Kinome showing major groups as well as families and subfamilies. Generated using ETE4 with data from ChEMBL.

2.2 Molecular Similarity

The similarity property principle (SPP), has been enormously influential in the realm of medicinal chemistry [15]. According to the SPP, structurally-similar compounds often exhibit similar properties. Among these properties is biological activity, such that similar compounds often demonstrate similar activity against identical biological targets (i.e., proteins in a majority of cases). It is important to note that the definition of "similar" here is ambiguous, and can be measured in a myriad of ways. Notable exceptions to the SPP are so-called "activity cliffs", where a compound which shows high binding affinity (i.e., high activity) against a given target becomes completely ineffective after minor structural modifications [15, 19].

There are many computational methods for computing molecular similarity between two compounds [14]. Similarity calculations depend both on how compounds are represented (e.g., using molecular graphs, fingerprint vectors, etc.) and the particular metric used. These similarity metrics typically report a value in the range 0 to 1, with values closer to 1 indicating a higher degree of structural similarity. In addition to computing pairwise similarity, dimensionality-reduction methods (e.g., U-MAP) can be used with clustering algorithms (e.g., K-means) to group together multiple structurally-similar compounds. Other clustering approaches make use of molecular scaffolds or other well-defined structural motifs.

For the purposes of this project, "similarity" is determined using Tanimoto coefficients computed from Morgan fingerprints generated by RDKit [11]. The family of Morgan fingerprints (also known as circular fingerprints or extended-connectivity fingerprints) are based upon an algorithm developed by H.L. Morgan [20]. The RDKit version of Morgan fingerprints is based upon the implementation described by Rogers and Hahn [23]. Although a detailed description of the Morgan fingerprint is outside the scope of this report, these fingerprints may be summarized as capturing the "presence of specific circular substructures around each atom in a molecule" [6]. Morgan fingerprints were chosen because they are widely used within the field of cheminformatics due to (1) the fact that their features are predictive of biological activity and (2) they have proven to be among the best performing fingerprints in virtual screening [6].

Like Morgan fingerprints, Tanimoto coefficients are widely adopted. The Tanimoto coefficient $T(A, B)$ of two bit vectors A and B is computed by the following equation:

$$T(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i - \sum_{i=1}^{n} a_i b_i} \tag{1}$$

Tanimoto values range from 0 to 1, with values closer to 1 indicating a higher degree of similarity. Figure 2 provides a visualization of some pairs of protein kinase ligands identified in this project and their corresponding Tanimoto similarity value.

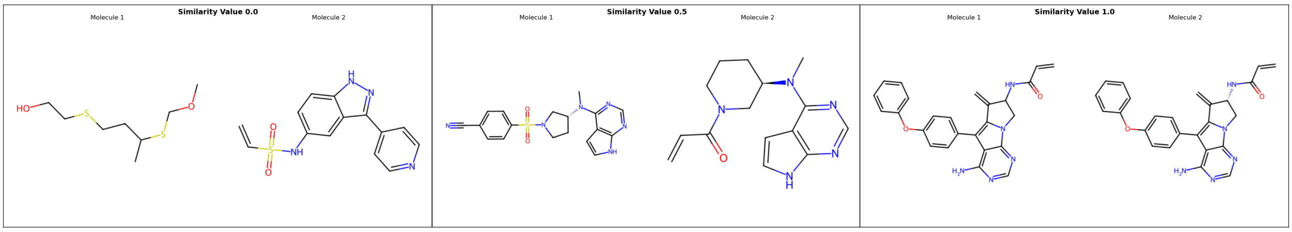

Figure 2: Pairs of protein kinase ligands and their Tanimoto similarity values (0, 0.5, 1).

2.3 Relevance to Drug Discovery

As mentioned above, protein kinases are involved in a number of human pathologies. If there is a relationship between protein kinase group and ligand similarity, then it may be informative to look at the ligands of related kinases (i.e., those belonging to the same group). This would be especially beneficial if there are few known ligands for the target of interest, as one could leverage data from well-studied, closely-related proteins to inform the discovery process.

2.4 Related Works

The idea that similar compounds may exhibit similar activities against proteins within the same family (or group) is not new. The term “intrafamily polypharmacology” (IFP) refers to molecules which have demonstrated activity against multiple proteins belonging to the same family, and has been of special interest in (for instance) studying PARP-1 inhibitors [22, 2]. Relatedly, previous works have investigated the theory of “Target-Family-Privileged Substructures”, which suggests that particular chemical substructures are strongly linked to activity against certain protein families [24]. Several works have developed computational analyses for looking at activity relationships between families [4], and works such as [19] and [13] have investigated the structures and activities of kinase inhibitors.

3 Dataset and Methodology

Within this project all data was sourced from ChEMBL (version 35) [26]. The ChEMBL PostgreSQL database was downloaded onto a local computer and was then used to carry out data extraction and analysis. A Snakemake [21] workflow was developed to extract relevant assays and ligands, and additional analyses were performed in Jupyter notebooks. The Snakemake workflow is shown in Figure 3, and will be explained in more detail below. As mentioned in the abstract, all code developed as part of this project can be found on GitHub: https://github.com/Jack-42/ligandActivityAnalysis.

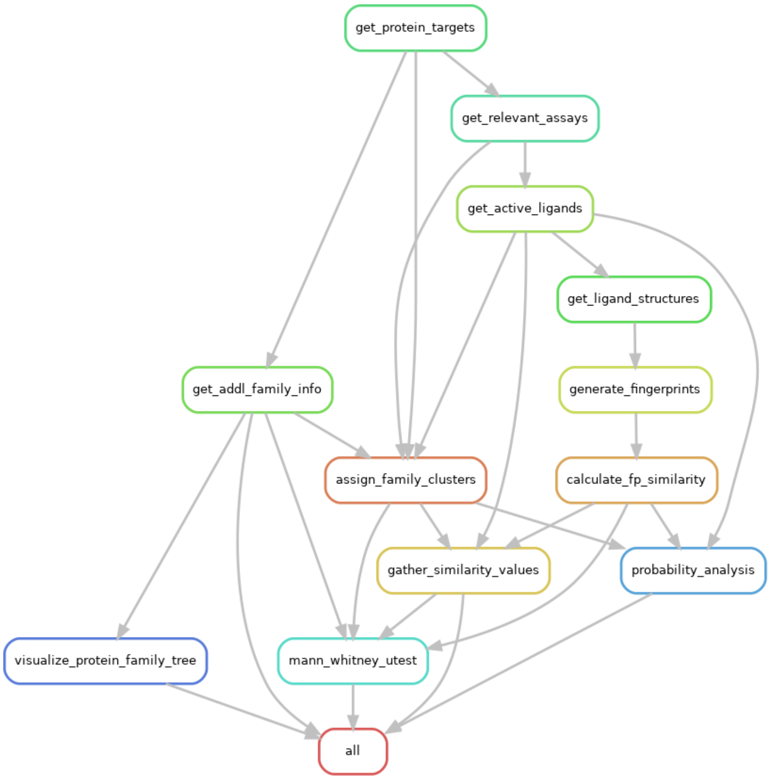

Figure 3: Snakemake workflow used to process and analyze ChEMBL data.

The first step of the workflow involved identifying all human protein kinases in ChEMBL. Here, any protein where the organism was equal to "Homo sapiens" ($ORGANISM$ = "Homo sapiens") and which belonged to the "Protein Kinase" class ($PROTEIN\_CLASS\_ID = 1100$) was selected. After identifying the set of protein targets, the set of active ligands and their relationship to specific human protein kinases was determined using data from single protein target binding assays linked to these targets in the ChEMBL database (version 35) [26]. In identifying appropriate assays and ligands, the criteria used by Pharos [10] was enforced to ensure that (1) assays were high-quality and (2) a reasonable definition of "active" was applied. These criteria include the following:

- Sample must have a pChEMBL value (i.e., a -Log M value)
- Must be from a binding assay
- Ligand must have a MOL structure type
- Assay must have a target type of $SINGLE\_PROTEIN$
- Sample must have $standard\_flag = 1$ and exact $standard\_relation$
- Assay must be associated with a journal publication
- Sample must have an activity value $\leq$ 30nM

In addition to the criteria from Pharos, the filters below were also applied:

- Removal of assays where target was a variant/mutant (using implementation described by [12]).
- Filtered out PAINS compounds to remove (some) false positives [3].
- Molecular weight of ligand must fall between [200, 900] Da (this particular range is based upon [19]).

Variable	Value
N. Protein Targets	423
N. Assays	73,487
N. Active Samples	38,622
N. Unique Ligands	9,995

Table 1: Counts of proteins, assays, samples, and unique ligands in the dataset

The application of the criteria above to the ChEMBL database (version 35) resulted in a dataset described in Table 1.

In terms of ligand-target relationships, these were directly inferred from the assays (i.e., if a ligand was present in an active sample in an assay where protein X was specified as the target, then the ligand was considered as active against protein X). As mentioned above, ChEMBL provides classification information for protein targets, and this information was used to assign ligands to respective kinase group(s) such that a ligand is considered active within a protein kinase group if they were identified as an active within an assay targeting a protein belonging to the group. It is perhaps worth noting that, under this definition, a single ligand can belong to more than one kinase group. Most commonly this is because the ligand has multitarget activity (i.e., it is active against more than one target, and these targets span more than one group)[1]. However, there are also some rare cases where proteins have been classified into more than a single group by ChEMBL. For example, Ribosomal protein S6 kinase alpha 2 belongs to both the AGC and CAMK group according to ChEMBL. In these exceptional cases an inclusive approach is taken such that the protein (and any of its active ligands) are considered part of both groups. A breakdown of the number of proteins, assays, and ligands belonging to each protein kinase group is provided in Figure 4.

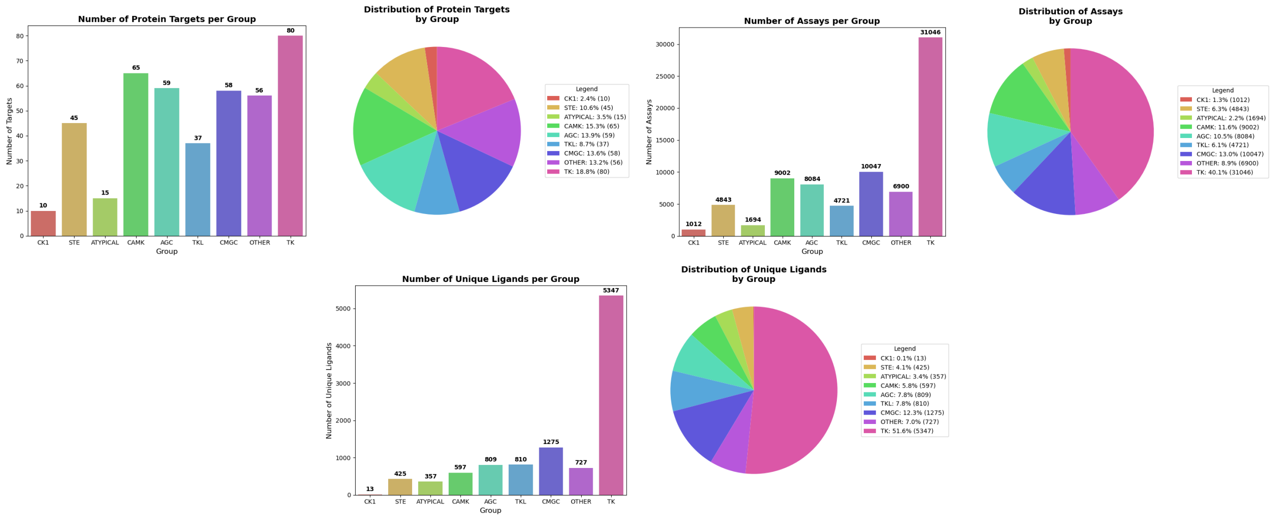

Figure 4: Number of unique protein targets (top left), assays (top right), and ligands (bottom) belonging to each protein kinase group in the dataset.

After determining the set of ligands and their group relationships, Morgan fingerprints were computed using RDKit. Tanimoto similarity coefficients were then computed between these 2D fingerprints. These generated similarity values were clustered together using the ligand-group relationships described above and to generate the results described in the next section.

[1] Within the dataset a total of $N = 8036$ (80.6%) of ligands were active against only a single target, $N = 1803$ (18.0%) were active against 2–4 targets, and $N = 156$ (1.6%) were active against ≥ 5 targets.

4 Results

To directly answer the main research question of this study (are ligands which are active within a particular protein kinase group more similar to one another than protein kinase ligands generally?), one can directly compare the distribution of similarity values computed with respect to all $9,995$ ligands in the dataset to the distribution of similarity values computed with respect to the ligands of individual protein kinase groups. These results are shown in Section 4.1. In addition, results show that there is (on average) a relationship between similarity threshold and enrichment with respect to both intra-group activity and assay activity (Section 4.2). However, as will be detailed in Section 5, there are several limitations to this study that must be accounted for before accepting the results presented below.

4.1 Distribution

Figure 5 provides distributions for the $\binom{N}{2}$ similarity values per group (N = number of ligands), as well as the $\binom{9{,}995}{2} = 49,945,015$ similarity values calculated for all kinase ligands in the dataset, with Table 2 providing the statistics of each of distribution.

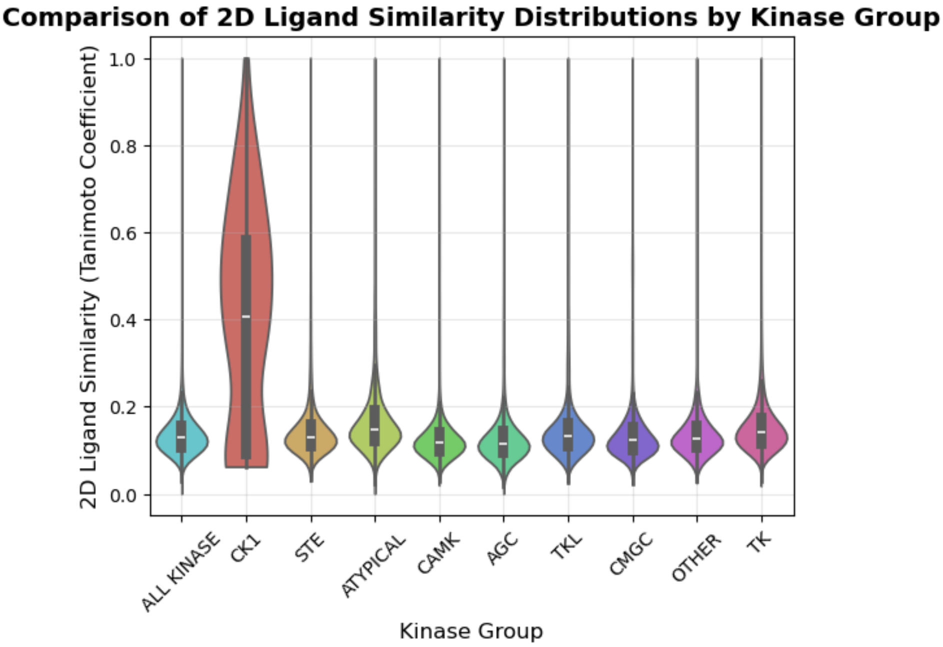

Figure 5: Comparison of 2D structural similarity distributions by kinase group. Due to computational limitations a maximum of 1,000,000 (randomly selected) similarity values are shown for each distribution (note this only impact the “ALL KINASE” and “TK” distributions). When performing statistical analyses the entire set of similarity values was always used.

Distribution	N. Similarity Values	Mean	Median	Variance	Skewness	Kurtosis
ALL KINASE	49,945,015	0.136	0.129	0.003	3.384	26.476
CK1	78	0.407	0.407	0.064	0.055	-1.016
STE	90,100	0.152	0.131	0.009	4.041	20.209
ATYPICAL	63,546	0.180	0.149	0.014	3.095	11.453
CAMK	177,906	0.133	0.117	0.006	4.895	29.897
AGC	326,836	0.135	0.116	0.009	4.158	21.775
TKL	327,645	0.153	0.133	0.009	3.693	17.759
CMGC	812,175	0.144	0.122	0.009	3.829	17.676
OTHER	263,901	0.144	0.127	0.007	4.441	25.778
TK	14,292,531	0.153	0.140	0.004	3.176	18.678

Table 2: Statistics for the distributions of ligand similarity values shown in Figure 5. Note that the number of similarity values belonging to each distribution will be equal to the number of pairs of ligands, i.e. $\binom{N}{2}$.

As can be seen from both Figure 5 and Table 2, with the exception of the CK1 group there is not a clear difference between the similarity of ligands active within a single protein kinase group compared to the similarity of all protein kinase ligands. To measure the statistical significance of these findings, a Mann-Whitney U test (MWUT) [16] was performed with respect to each of the group distributions (note that since we have a total of 9 "groups" we are performing 9 different tests). The MWUT was chosen as an appropriate statistical test because of the following considerations:

1. It is a non-parametric test (note that the distributions are not clearly normal, exponential, etc.).
2. The MWUT tests whether two samples have the same underlying distribution. Thus, the test is more informative than just comparing a single statistic such as the mean or median.

The test was performed using the implementation provided by SciPy [25, 1]. If we let X_G be the distribution of similarity values computed with respect to all ligands active within a given protein kinase group G (CK1, STE, etc.), and Y be the distribution of similarity values computed with respect to all other protein kinase ligands in the dataset[2], then the test has the following null and alternative hypotheses [1]:

> H_0: The distribution underlying X_G is not stochastically greater than the distribution underlying Y.
>
> H_a: The distribution underlying X_G is stochastically greater than the distribution underlying Y.

The test reports both a p-value and a U statistic. Since we're running multiple tests on different distributions, a Bonferonni correction is applied to all p-values to account for the multiple comparisons problem. The U statistic is considered the maximum value of $U1$ and $U2$ below:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{2}$$

Where n_1 and n_2 are the number of observations (i.e., similarity values) in X_G and Y respectively, and R_1 and R_2 are the sums of the ranks in X_G and Y after ranking all values such that the smallest value obtains rank 1 and the largest rank $n_1 + n_2$.

The results from performing the MWUT are shown in Table 3. As can be seen from the p-values in the table, according to the MWUT there is reason to reject the null hypothesis (H_0) stated above for several of the protein kinase groups. However, except for the CK1 group, the effect size is not very large (the extreme p-values are a result of the very large magnitude of n_1 and n_2). In the case

[2] Note that this definition of Y is not the same as the "ALL KINASE" distribution shown above, as all similarity values computed with respect to ligands active in G are removed. This is necessary to ensure that X_G and Y are independent.

of the CK1 group, both a large effect size and significant p-value were observed. However, it should be noted that the number of ligands within the CK1 group is only 13, which in the wider context of chemistry + biology is unlikely to be a representative sample of all molecules which bind to CK1 proteins. There are also additional limitations of this study, such as the fact that not all proteins are equally well-represented in the dataset, that prevents these results from being definitive. These limitations will be discussed in more detail in Section 5.

Group	Group size (n_1)	Comparison size (n_2)	Group Median	Comparison Median	U_1	U_2	p-value
CK1	78	49,815,171	0.407	0.129	1.4002×10^{13}	1.6873×10^{13}	3.10×10^{-11}
STE	90,1000	45,787,665	0.131	0.129	2.1731×10^{12}	1.9524×10^{12}	2.48×10^{-169}
ATYPICAL	63,546	46,440,703	0.149	0.129	1.8801×10^{12}	1.0710×10^{12}	$< 5 \times 10^{-324}$
CAMK	17,7906	44,156,503	0.117	0.130	3.1847×10^{12}	4.6710×10^{12}	1.0
AGC	32,6836	42,186,705	0.116	0.131	5.4708×10^{12}	8.3173×10^{12}	1.0
TKL	327,645	42,177,520	0.133	0.129	7.3358×10^{12}	6.4834×10^{12}	$< 5 \times 10^{-324}$
CMGC	81,2175	38,014,840	0.122	0.130	1.4002×10^{13}	1.6873×10^{13}	1.0
OTHER	26,3901	42,943,278	0.127	0.129	5.5731×10^{12}	5.7596×10^{12}	1.0
TK	14,292,531	10,799,628	0.140	0.121	9.9533×10^{13}	5.4821×10^{13}	$< 5 \times 10^{-324}$

Table 3: Comparison of similarity distributions computed with respect to ligands active within particular protein kinase groups (X_G) vs the similarity distribution computed with respect to all other ligands in the dataset (Y). Shown p-values are calculated (with Bonferroni correction) from a Mann-Whitney U test (MWUT) where the alternative hypothesis is that the similarity values within the group are stochastically greater than the distribution of all similarity values. Note that although the MWUT is *not* a test which considers the difference of medians, the median values are reported here to provide the reader with a more intuitive/familiar metric than the U statistic.

4.2 Enrichment

In addition to looking at the overall distribution of similarity values, the enrichment factor per group was also investigated. Enrichment is defined here as how much more likely it is (on average) that two ligands are active within the same group G given that their similarity is higher than some threshold. Prior research has looked at versions of this question before, and has found that when molecules are highly similar (Tanimoto similarity $\geq$ 0.85) there is a $20 - 30\%$ chance they have similar biological activity [18, 7]. These probabilities are higher than random chance, but certainly leave room for exceptions. In the context of this work "enrichment" is defined according to equation 3:

$$\frac{P(\text{two ligands active within group G} \mid \text{similarity} > \text{threshold})}{P(\text{two ligands active within group G})} \tag{3}$$

The motivation behind looking at enrichment is that, while there may not be a significant difference in the overall distributions of similarity values, one may still observe a higher-than-chance probability that two *highly similar* ligands are active in the same group, and thus there may be a relationship between ligand similarity and intra-group activity at a sufficient threshold.

Figure 6 shows the enrichment values observed for different protein kinase groups. As is seen from the figure, there is generally a positive correlation between similarity threshold and the enrichment factor. However, there are good reasons to be skeptical of these results. It is unexpected for there to be a strong relationship between similarity and enrichment within either the "Atypical" or "Other" group, since the proteins within these groups are not related to one another in a meaningful way. Additionally, no relationship was observed for the TK group, even though these proteins are related in a meaningful way (they almost exclusively phosphorylate tyrosine residues). These counter-intuitive results are likely due to an imperfect definition of enrichment being used, as well as the general study limitations described in Section 5.

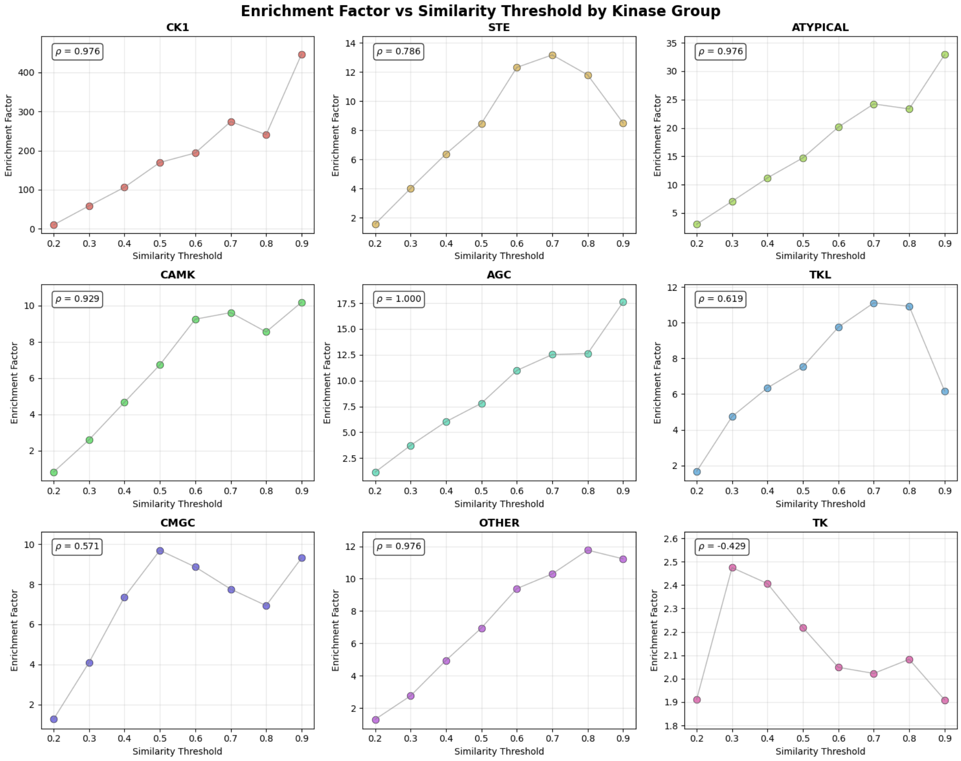

Figure 6: Plot of Tanimoto similarity threshold (x-axis) vs enrichment (y-axis) by protein kinase group.

The first issue with the definition of enrichment given above is that the magnitude of the enrichment factor depends upon the baseline probability of P(two ligands active within group G). While it intuitively makes sense to normalize the conditional probability by the baseline when looking at enrichment, in practice this results in the enrichment factor being inflated when the baseline probability is extremely low. If one compares the baselines shown in Figure 7 to the observed enrichment values in Figure 6, they will see that there is a correlation between the baseline probability and observed enrichment values.

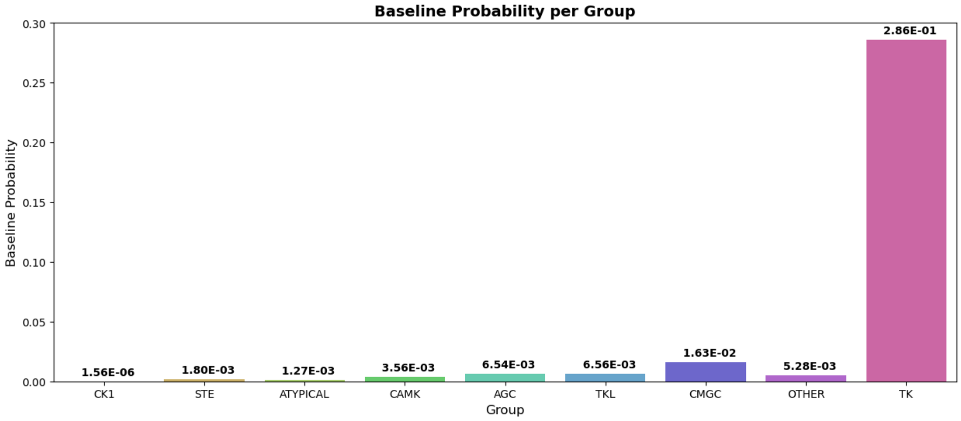

Figure 7: Baseline probability P(two ligands active within group G) for each protein kinase group. Note that the probabilities here will not sum to 1 because the sum of these baselines merely represents the probability that two randomly chosen ligands are both active within the same group.

Another limitation is that the enrichment definition above does not account for ligands being active against the same target or within the same assay. Thus, the enrichment factor could be high for a given group simply because highly similar ligands are more likely to be active against the same target or within the same assay, but not necessarily because they are more likely to be active against other proteins in the same group. Figure 8 shows that, while there is a significant spread in enrichment factor between individual assays, on average the enrichment factor exceeds 100 when the similarity threshold is greater than 0.5, and is close to 1000 when the similarity threshold is greater than 0.8. Since the per-assay enrichment factor is so enormous (on average), it must be taken into account when considering the results above.

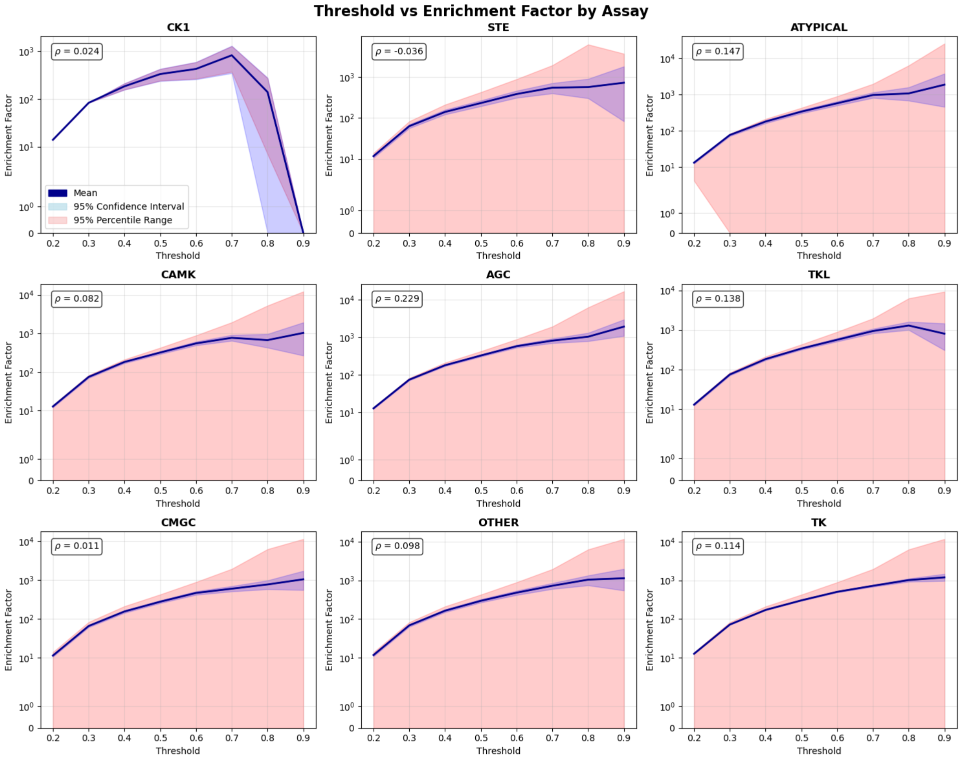

Figure 8: Plot of Tanimoto similarity threshold (x-axis) and enrichment factor (y-axis) by assay. Here only assays which had at least 2 active ligands were considered. Assays were clustered based on the group membership of targets. It is worth noting that the enrichment factor is equal to 0 when none of the active ligands in the assay have a similarity value greater than the threshold, which is why such a large percentile range is observed in each plot. It is also why there is a notable drop in the mean enrichment factor at a threshold of 0.8 for CK1 assays.

5 Discussion

As has been mentioned above, there are several limitations to this study which prevent the results presented in the previous section from being definitive. The first of these limitations is that not all protein targets within the dataset are equally well-represented. As is shown by Figure 9, some proteins in the dataset are tested in both a much larger number of assays and (often as a result) have a much larger number of active ligands than other proteins – including those within the same group. There are several proteins which were not tested in any of the selected 73, 487 assays, and thus were not at all represented in the dataset. While it makes sense that some proteins would not be as well-researched as others (e.g., because they are not known to be relevant to a prominent disease), in the context of this project this fact is problematic because it means the results above are biased towards proteins with a large number of known active ligands. An improved study design could try to account for this fact by considering many samples of the 9, 995 ligands, such that within each sample an equal number of ligands per protein target were considered. However, such an approach would still be unable to account for the proteins which have 0 active ligands.

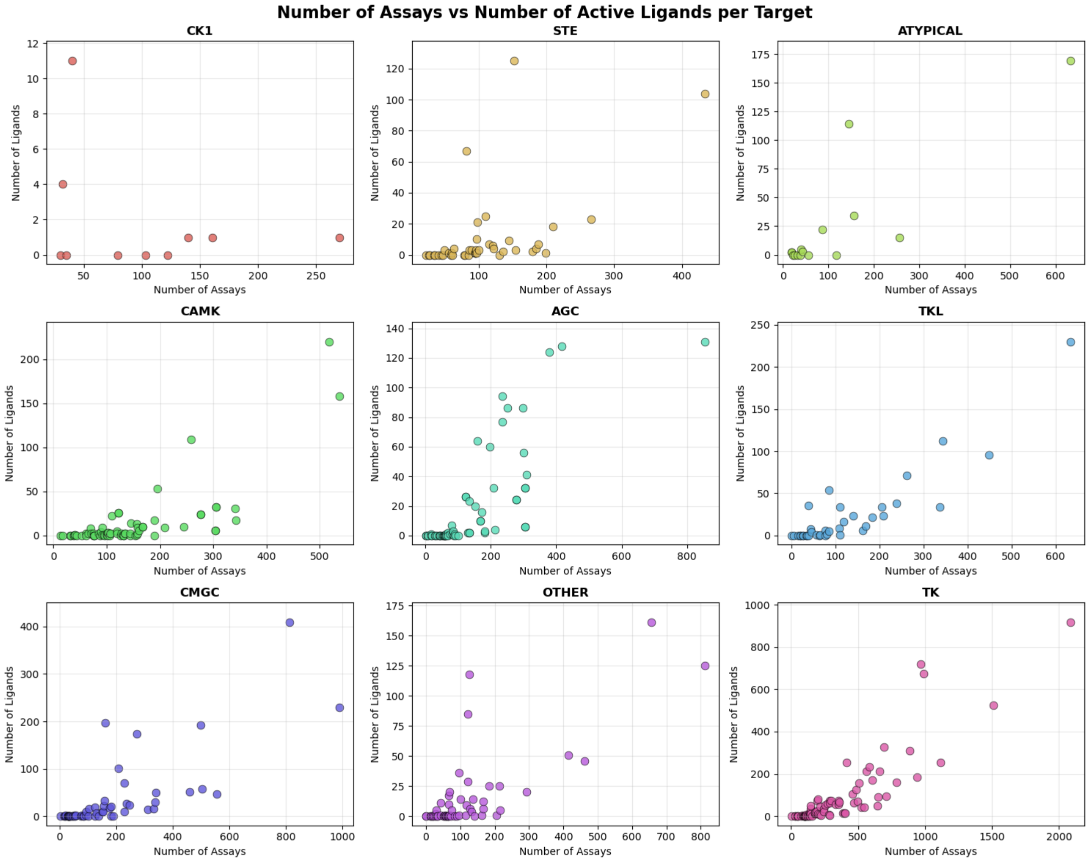

Figure 9: Scatter plots showing the number of assays (x-axis) and ligands (y-axis) for each protein belonging to each protein kinase group.

Another limitation of the dataset is that there is immense amount of noise in assay data, and an even greater amount of noise when combining data from a large number of assays from ChEMBL [12]. While the dataset used in this project was curated using criteria developed by other researchers, there is still the difficult fact that the methodology does not directly account for differences in conditions between assays. Among these conditions is what compounds were actually tested in each assay. The 9, 995 ligands in our dataset were not all tested against every protein target in the dataset, and thus there are likely many cases where a ligand was not assigned to a kinase group simply because it hadn't been tested against many of the targets in that group (if any).

In addition to limitations of the dataset, it is worth noting also that this project only considered a single definition of ligand structural similarity. While Morgan fingerprints and Tanimoto coefficients are a popular means of computing similarity, they are certainly not the only option [15, 14].

Finally, there are chemical and biological realities that help explain the lack of significant results from this study. For one, a single protein can have multiple binding sites, and one should not necessarily expect two ligands which bind to the same protein to be "similar" to one another. The methodology implicitly assumes that PKs in the same group will have more structurally similar binding sites compared to other PKs (due to their evolutionary conservation, shared functions, etc.), but perhaps this assumption should be tested. Another fact is that even highly similar compounds can have significant differences in potency against the same target (activity cliffs). As noted in [18], designing compound libraries based on "similarity" is a matter of probability.

6 Conclusion

Overall, this study found no clear relationship between 2D ligand similarity and protein kinase group activity. Although significant differences were observed for the CK1 group, given the limitations outlined above further study is necessary before accepting these results.

Despite the lack of positive results from this study, it is worth noting that another outcome of this work was building a pipeline to process large amounts of data from ChEMBL. The Snakemake workflow built for this project (Figure 3) could likely be adapted to improve on not only this work, but also for use in other projects relying on ChEMBL data.

7 Acknowledgements

Much thanks to my advisor Dr. Vincent Metzger for his guidance and input over the course of this project. I'd also like to thank Dr. Jeremy Yang, Dr. Cristian Bologa, and Dr. Praveen Kumar for their feedback during weekly meetings over the course of the internship. Finally, I'd like to thank the authors of ChEMBL [26] for organizing such large amounts of assay data and for making ChEMBL free and publicly accessible.

References

[1] https://docs.scipy.org/doc/scipy-1.16.0/reference/generated/scipy.stats.mannwhitneyu.html, 2025.

[2] Albert A. Antolin, Malaka Ameratunga, Udai Banerji, Paul A. Clarke, Paul Workman, and Bissan Al-Lazikani. The kinase polypharmacology landscape of clinical parp inhibitors. *Scientific Reports*, 10(1), Feb 2020.

[3] Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, Apr 2010.

[4] Jürgen Bajorath. Computational analysis of ligand relationships within target families. *Current Opinion in Chemical Biology*, 12(3):352–358, Jun 2008.

[5] Khushwant S. Bhullar, Naiara Orrego Lagarón, Eileen M. McGowan, Indu Parmar, Amitabh Jha, Basil P. Hubbard, and H. P. Vasantha Rupasinghe. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular Cancer*, 17(1), Feb 2018.

[6] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1), Jun 2020.

[7] B Chen, P Greenside, H Paik, M Sirota, D Hadley, and AJ Butte. Relating chemical structure to cellular response: An integrative analysis of gene expression, bioactivity, and structural data across 11,000 compounds. *CPT: Pharmacometrics and Systems Pharmacology*, 4(10):576–584, Sep 2015.

[8] Lubos Cipak. Protein kinases: Function, substrates, and implication in diseases. *International Journal of Molecular Sciences*, 23(7):3560–3560, Mar 2022.

[9] Sameh Eid, Samo Turk, Andrea Volkamer, Friedrich Rippmann, and Simone Fulle. Kinmap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics*, 18(1), Jan 2017.

[10] Keith J Kelleher, Timothy K Sheils, Stephen L Mathias, Jeremy J Yang, Vincent T Metzger, Vishal B Siramshetty, Dac-Trung Nguyen, Lars Juhl Jensen, Dušica Vidović, Stephan C Schürer, Jayme Holmes, Karlie R Sharma, Ajay Pillai, Cristian G Bologa, Jeremy S Edwards, Ewy A Mathé, and Tudor I Oprea. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Research*, 51(D1), Nov 2022.

[11] Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, guillaume godin, Axel Pahl, tadhurst cdd, Juuso Lehtivarjo, Francois Berenger, and jasondbiggs. rdkit/rdkit: 2024_03_3 (Q1 2024) Release. https://doi.org/10.5281/zenodo.11396708, May 2024.

[12] Gregory A Landrum and Sereina Riniker. Combining ic50 or ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5), Feb 2024.

[13] Yu-Chen Lo, Tianyun Liu, Kari M Morrissey, Satoko Kakiuchi-Kiyota, Adam R Johnson, Fabio Broccatelli, Yu Zhong, Amita Joshi, and Russ B Altman. Computational analysis of kinase inhibitor selectivity using structural knowledge. *Bioinformatics*, 35(2):235–242, Jul 2018.

[14] Kenneth López-Pérez, Juan F Avellaneda-Tamayo, Lexin Chen, Edgar López-López, K. Eurídice Juárez-Mercado, José L Medina-Franco, and Ramón Alain Miranda-Quintana. Molecular similarity: Theory, applications, and perspectives. *Artificial Intelligence Chemistry*, 2(2):100077–100077, Aug 2024.

[15] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, Nov 2013.

[16] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, Mar 1947.

[17] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[18] Yvonne C. Martin, James L. Kofron, and Linda M. Traphagen. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, Sep 2002.

[19] Filip Miljković and Jürgen Bajorath. Computational analysis of kinase inhibitors identifies promiscuity cliffs across the human kinome. *ACS Omega*, 3(12):17295–17308, Dec 2018.

[20] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965.

[21] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. *F1000Research*, 10:33, Jan 2021.

[22] Daniela Passeri, Emidio Camaioni, Paride Liscio, Paola Sabbatini, Martina Ferri, Andrea Carotti, Nicola Giacchè, Roberto Pellicciari, Antimo Gioiello, and Antonio Macchiarulo. Concepts and molecular aspects in the polypharmacology of parp-1 inhibitors. *ChemMedChem*, 11(12):1219–1226, Oct 2015.

[23] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, Apr 2010.

[24] Dora M. Schnur, Mark A. Hermsmeier, and Andrew J. Tebben. Are target-family-privileged substructures truly privileged? *Journal of Medicinal Chemistry*, 49(6):2000–2009, Mar 2006.

[25] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and C J Carey. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, Feb 2020.

[26] Barbara Zdrazil, Eloy Félix, Fiona Hunter, Emma Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Méndez, Juan F Mosquera, María Paula Magariños, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A. Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1), Nov 2023.