Analyzing Patterns of Computational Similarity between Protein Kinase Ligands

Jack Ringer

July 2025

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 kinase ligands was compared against the distribution of pairwise similarity values computed with respect to ligands active within a specific kinase group. With the exception of 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 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.

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. 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.

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 phosophorylation [3]. PKs are involved in many important regulatory roles throughout the cell, and their dyregulation is linked to many types of cancer as well as immune, neurological and infectious diseases [**?**, **?**]. Although all PKs perform phosophorylation, 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 [7]. 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 [4]. 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 exception to this is the "Atypical" and "Other" groups, which serve to classify proteins which don't fit into the other major groups. A phylogenetic tree of the human kinome is shown in Figure 1.

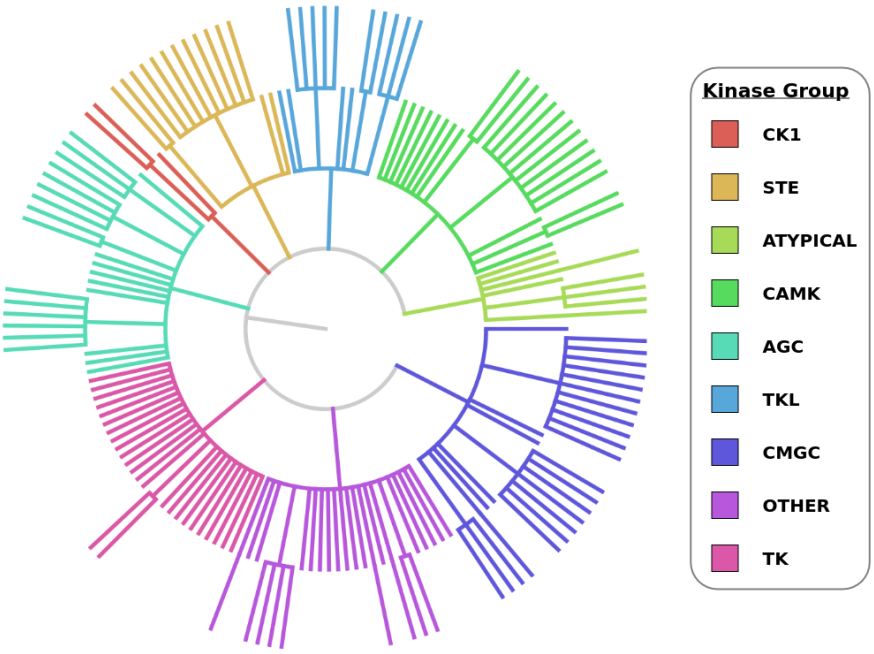

Figure 1: Phylogenetic Tree of the Human Kinome showing major groups as well as families and subfamilies. Generated using ETE4 with data from CheMBL.

Molecular Similarity

The similarity property principle (SPP), has been enormously influential in the realm of medicinal chemistry [6]. 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. Additionally, there exist 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 [6, 8].

There are many computational methods for computing molecular similarity between two compounds [4]. Similarity calculations depend both on how compounds are represented (e.g., using molecular graphs, fingerprint vectors, etc), as well as 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...

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.

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 [9, 1]. 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 [10]. Several works have developed computational analyses for looking at activity relationships between families [2], and works such as [8] and [5] have investigated the structures and activities of kinase inhibitors.

Methodology

Results

Distribution

Figure 2 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 calcualted for all kinase ligands in the dataset. Table 1 provides additional statistics for each group, as well as results from a Mann-Whitney U test (MWUT).

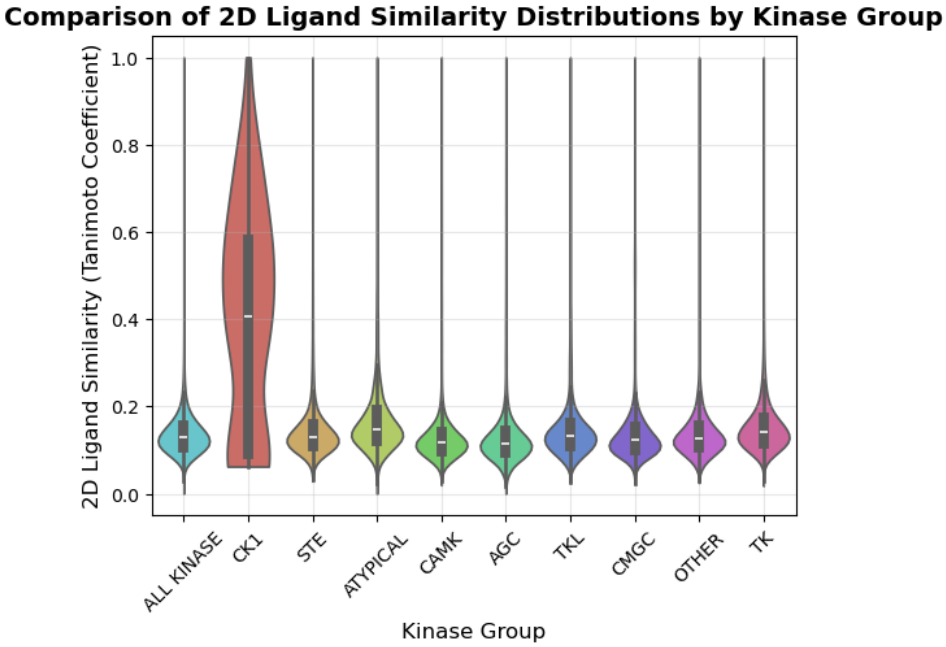

Figure 2: Comparison of 2D structural similarity distributions by kinase group

Kinase Group	N. Targets	N. Ligands	Group Median	Comparison Median	MWUT p-value
CK1	10	13	0.407	0.129	3.10×10^{-11}
STE	45	425	0.131	0.129	2.48×10^{-169}
ATYPICAL	15	357	0.149	0.129	$< 5 \times 10^{-324}$
CAMK	65	597	0.117	0.130	1.0
AGC	59	809	0.116	0.131	1.0
TKL	37	810	0.133	0.129	$< 5 \times 10^{-324}$
CMGC	58	1275	0.122	0.130	1.0
OTHER	56	727	0.127	0.129	1.0
TK	80	5347	0.140	0.121	$< 5 \times 10^{-324}$

Table 1: Table showing comparisons of targets, ligands, and ligand similarity distributions per kinase group. Shown p-values are calculated (with Bonferroni correction) from a MWUT where the alternative hypothesis is that the similarity values within the group are stochastically greater than the distribution of all similarity values.

Enrichment

Discussion

Conclusion

Acknowledgement

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 DB [11].

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.