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

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 [6]. 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 [4]. 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 [14]. 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 [7]. 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.

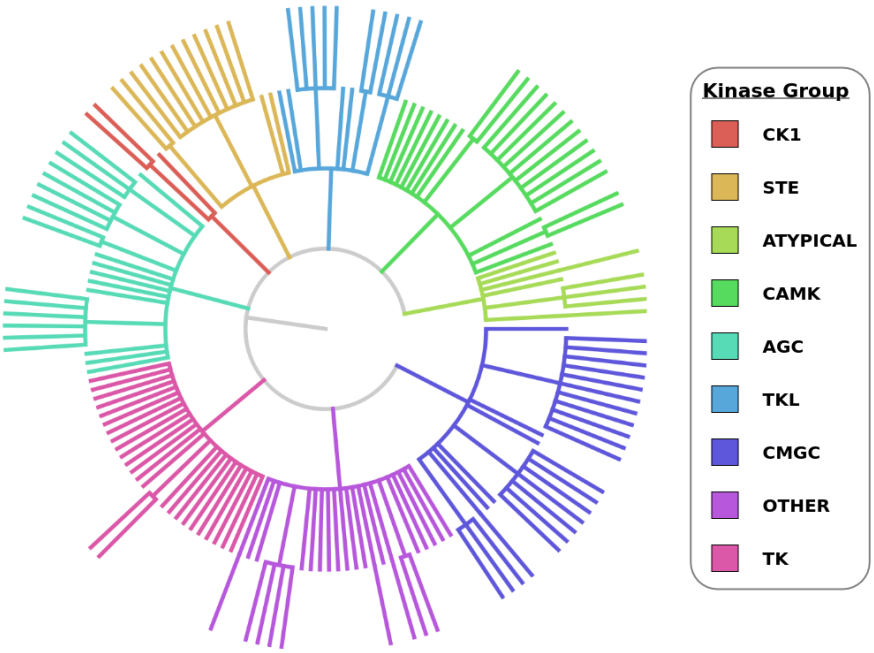

Figure 1: Phylogenetic Tree of the Human Kinome showing major groups as well as families and subfamilies. Generated using ETE4 with data from ChEMBL.

Molecular Similarity

The similarity property principle (SPP), has been enormously influential in the realm of medicinal chemistry [13]. 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 [13, 15].

There are many computational methods for computing molecular similarity between two compounds [12]. 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, "similarity" is determined using Tanimoto coefficients computed from Morgan fingerprints generated by RDKit [9]. The family of Morgan fingerprints (also known as circular fingerprints or extended-connectivity fingerprints) are based upon an algorithm developed by H.L. Morgan [16]. The RDKit version of Morgan fingerprints is based upon the implementation described by Rogers and Hahn [19]. 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" [5]. 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 [5].

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.

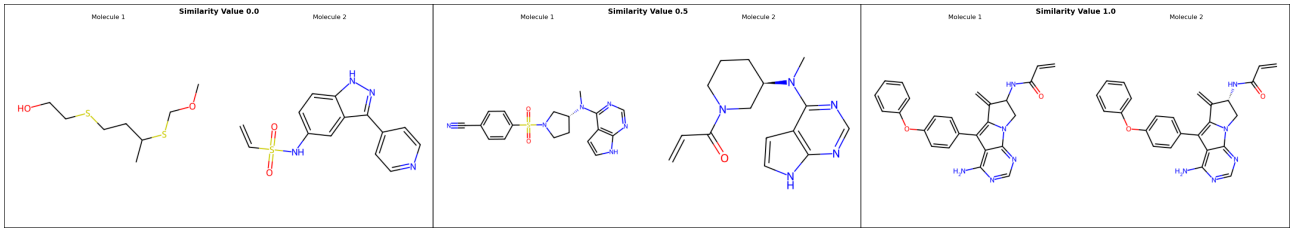

Figure 2: Pairs of protein kinase ligands and their Tanimoto similarity values (0, 0.5, 1).

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 [18, 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 [20]. Several works have developed computational analyses for looking at activity relationships between families [3], and works such as [15] and [11] have investigated the structures and activities of kinase inhibitors.

Dataset and Methodology

Within this project all data was sourced from ChEMBL (version 35) [21]. The ChEMBL PostgreSQL database (DB) was downloaded onto a local computer and was then used to carry out data extraction and analysis. A Snakemake [17] workflow was developed to extract relevant assays and ligands (see details below), and additional analyses were performed in Jupyter notebooks. As mentioned in the abstract, all code developed as part of this project can be found on GitHub: https://github.com/Jack-42/ligandActivityAnalysis.

The set of active ligands and their relationship to specific human protein kinases was determined using data from single protein target binding assays in the ChEMBL database (version 35) [21]. In identifying appropriate assays and ligands, the criteria used by Pharos [8] 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, other filters were also applied based on other prior research and consideration. In particular:

- Remove assays where target was a variant/mutant (using implementation described by [10]).
- Filtered out PAINS compounds to remove (some) false positives [2].
- Molecular weight of ligand must fall between [200, 900] Da (this particular range is based upon [15]).

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

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

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). 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 3.

[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.

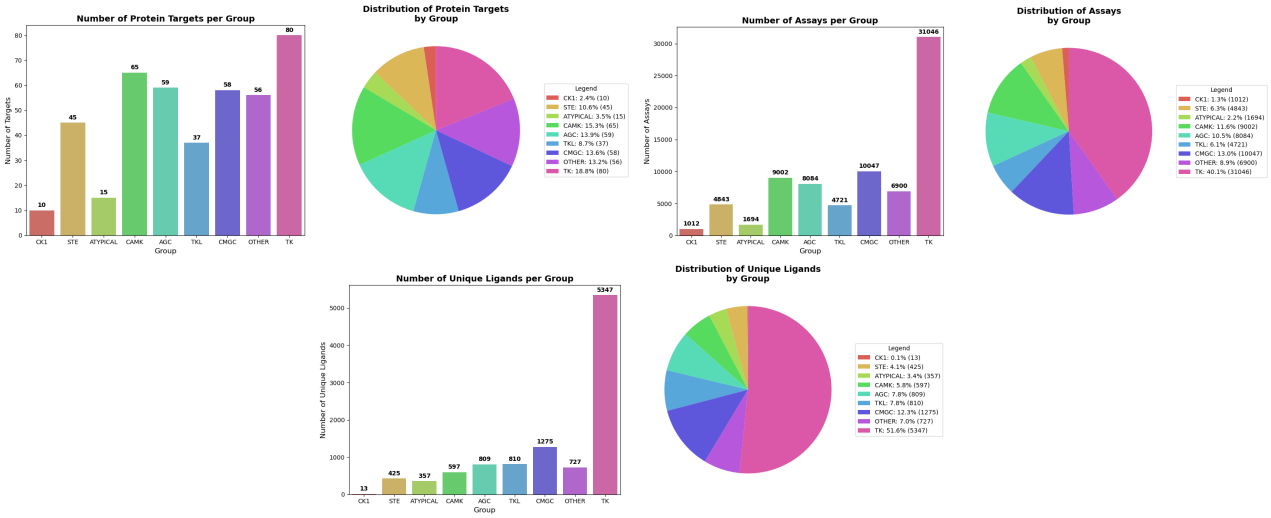

Figure 3: Number of unique protein targets (top left), assays (top right), and ligands (bottom) belonging to each protein kinase group in the dataset.

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.

Results

Distribution

To directly answer. Figure 4 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 2 provides additional statistics for each group, as well as results from a Mann-Whitney U test (MWUT).

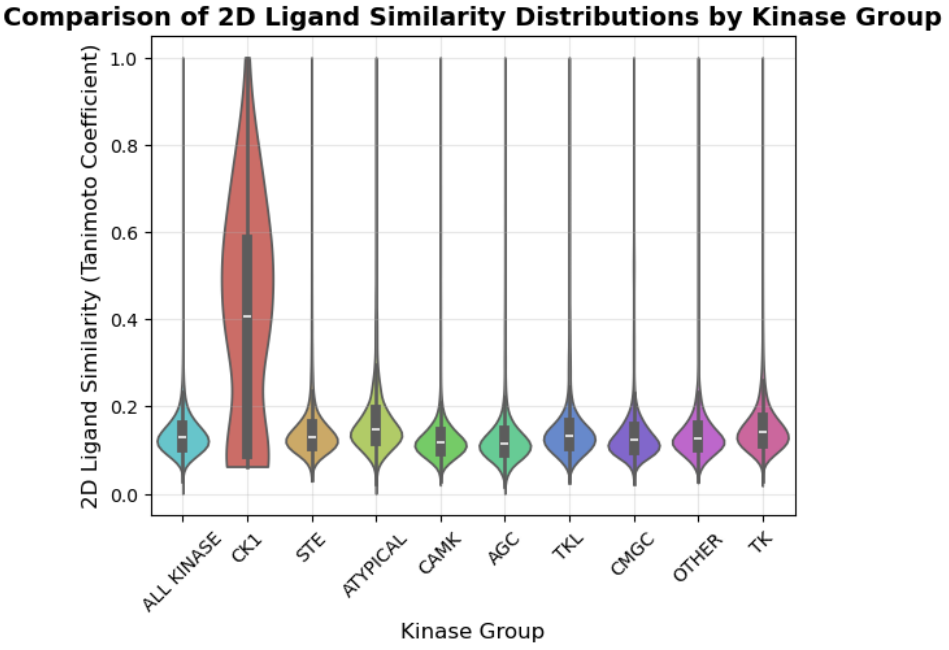

Figure 4: Comparison of 2D structural similarity distributions by kinase group

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 2: 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 [21].

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] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.