

## Introduction

Protein kinases are enzymes that are involved in several cellular pathways. These enzymes catalyse the transfer of a phosphoryl group from ATP to an acceptor which can either be protein substrates, lipids or small molecules. Protein kinases catalyse the transfer of  $\gamma$ -phosphate of ATP to the hydroxyl groups of various substrates. Through this phosphorylation process, they could modify enzymatic activities which leads to alteration in biological processes such as the control of metabolism, transcription processes, cell division and movement, programmed cell death and several other signal transduction events in the cell.

Protein kinases are the second largest enzyme family and the fifth largest family of genes in human following zinc finger proteins, G-protein coupled receptors, immunoglobulins, and the protease enzymes [?]. About 2% of the entire genome have been shown to encode protein kinases. [?] in 2002, identified all sequenced eukaryotic protein kinases by searching every available human genome sequence source such as Celera genomic databases, Incytes EST, Genbank cDNAs and expressed sequence tags (ESTs) using hidden Markov Model (HMM) profiling of the sequences to profile the protein kinase domains [?]. Overall, they identified 518 human protein kinase genes in which 478 were classified as eukaryotic protein kinases (ePKs) while 40 were Atypical protein kinases (aPK) which lack sequence similarity to the eukaryotic kinase domain but have been reported to have biochemical kinase activity.

## Protein Kinase Superfamily

The catalytic domain and the kinase domain of eukaryotic proteins are highly conserved both in sequence and structure. Protein kinase activity requires the binding of a peptide substrate—which is to be phosphorylated—and the Mg-ATP to the catalytic domain. Protein kinases can be broadly classified as either tyrosine kinases or serine/threonine kinases based on the specificity of the substrate they phosphorylate. Kinases are divided into groups, families as well as subfamilies. The protein kinases are classified into 9 groups based on the sequence and structural similarities of the catalytic domain. The classification was aided by knowledge of the sequence similarity and domain structure outside the catalytic domains, known biological functions and evolutionary history of the kinases. This classification is an extension of the work by Hanks and Hunter who initially performed

a conservation and phylogeny analysis of the catalytic domain of eukaryotic protein to reveal the conserved features of catalytic domains and thus, classified the protein kinase into 5 groups, 44 families and 51 subfamilies [?] while [?] further extended this to 9 groups, 134 families and 196 subfamilies. Figure 1 below shows the grouping of the human protein kinases.

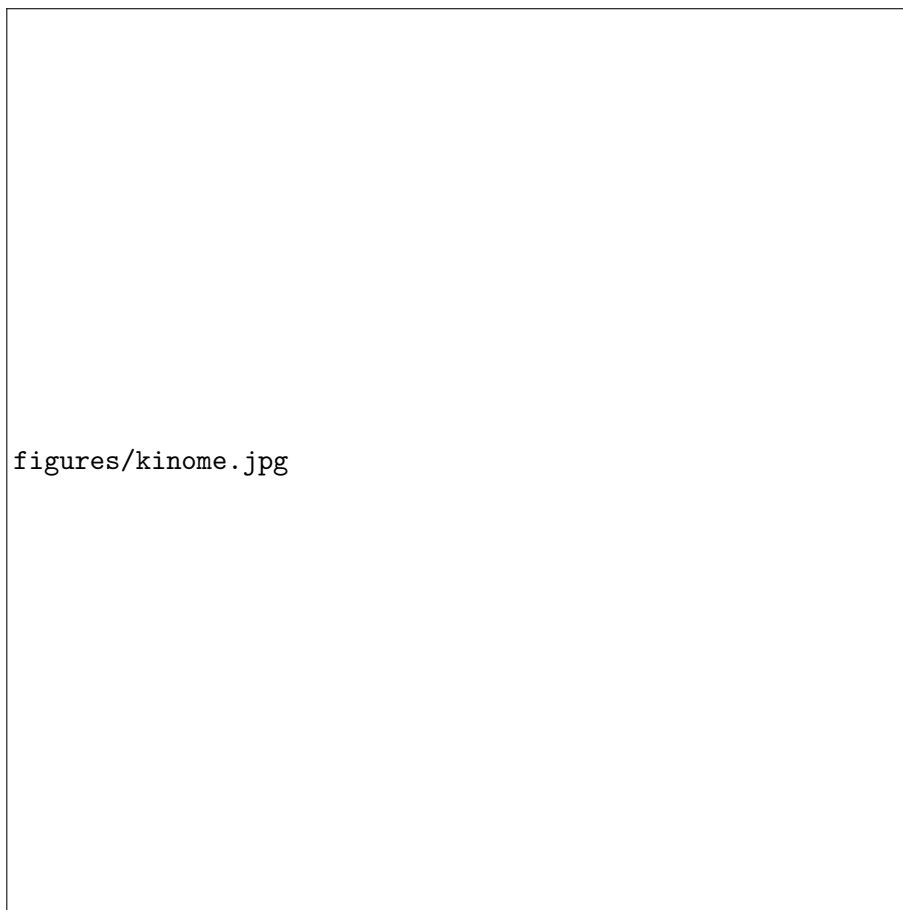


Figure 1: The human kinome. Kinome illustration courtesy of Cell Signalling Technology, Inc ([www.cellsignal.com](http://www.cellsignal.com)) based on [?].

## **Classification of Protein Kinases**

Below, the groups identified by [?] are described in more detail.

### **AGC Kinases**

These group of kinases include PKA, PKG, and PKC. They are involved in diverse cellular roles such as cell growth and proliferation, cell survival, glucose metabolism and protein

synthesis. They are also dysregulated in several diseases such as cancer and neurological disorder, inflammation and viral infection [?]. The Akt isoform possesses the pleckstrein homology domain (PH-domain) at the N-terminal and with this domain, it interacts with PIP3 and PIP2 which leads to the activation of pyruvate dehydrogenase kinase isoenzyme (PDK1). PKC also interacts with DAG and calcium by its N-terminal conserved domains (C1 and C2) which leads to conformational changes and activation of the protein [?].



Figure 2: The domain structure of AGC kinase family. All members contain Thr/Ser in the activation loop. Figure taken from [?]

### CAMK Kinases

These kinases are involved in calcium signalling and are basically autoinhibited. The binding of  $\text{Ca}^{2+}$ /calmodulin complexes relieves this autoinhibition. Members of this group include MLCK, RAD53, PKD, CAMK2, Trio, CAMKL, DCAMKL, CASK, and DAPK sub-families and they are all in multidomain architectures. Each member of this family possess additional unique domains in addition to the conserved kinase domain. For instance, the CASK ( $\text{Ca}^{2+}$  /calmodulin dependent serine kinases) contains a series of L27 repeats,

PDZ, two Src homology 3 (SH3) domain as well as a C-terminal guanylate kinase domain [?]. The PKD kinase also possesses a PH domain as found in the Akt family and this is important for the regulation of its enzymatic activity.

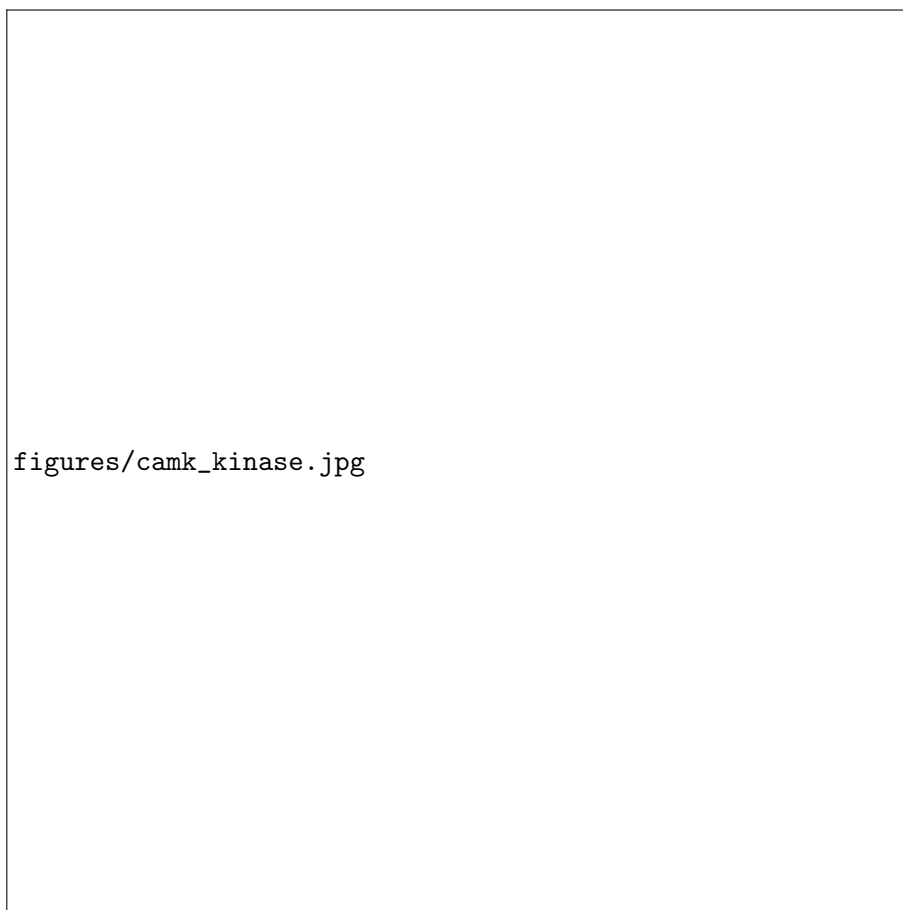


Figure 3: Domain organisation and structure of CaMKII. (A) Similar domain organisation in the CaMKII $\alpha$  and CaMKII $\beta$  with the exception of F-actin binding domain inserted in CaMKII $\beta$ . (B) Structure of autoinhibited CaMKII subunit PDB ID: 3SOA, (C) Cartoon showing the compact inactive holoenzyme (D) Cartoon showing conformational changes associated with CaMKII activation.

### CK1 group

The members of this group are quite ubiquitous in their phosphorylation events as they have a wide range of substrates. They are Ser/Thr kinases and are constitutively expressed. The kinases in this group do not possess additional non-catalytic domains apart from the CK1-gamma subfamily which possesses CK1-gamma domain whose function is not yet known.



Figure 4: Domain structure of human CK1 $\delta$ y. The members share a common conserved kinase domain but differs in their variable N-and-C terminal domains. The regulatory c-terminal domain has multiple inhibitory autophosphorylation sites. The nuclear localization signal(NLS) and kinesin homology domain (KHD) are also located within the kinase domain. Figure obtained from [?]

### CMGC group

Members of this group possess single domains like the CK1 group. They include dual specificity tyrosine regulated kinases, dual specificity yak-related kinases(DRYK), cyclin-dependent kinases (CDKs), MAPK, GSK-3, CDK-like kinases. CDKs regulates the progression through the different phases of the cell cycle in association with their activating partners cyclins. The MAP kinases are amongst the highly studied signal molecules. The MAP kinase cascade include control of proliferation, differentiation, and cell-death across various eukaryotes. The GSK-3 kinases are key metabolic enzyme in glycogen metabolism and play a role in the *Wnt* pathway which is important in embryonic development.

**Tyrosine Kinase group (TK-group)**

These kinases catalyse the phosphorylation of tyrosine residues and are heavily implicated in cancer. They are classified into receptor and non-receptor (cytosolic) kinases. The receptor TKs are classified based on the sequence homology and the structure of their extracellular domains into 20 families. One of the most studied extracellular domain is the Ig-like domain which occurs in most of the members of this subgroup. The extracellular domains acts as the ligand binding sites for several growth receptors. The non-receptor kinases are subdivided into 10 families which include Src, Abl, Ack, Csk, Fak, Fes, Frk/Fyn, Tec and Syk [?]. In addition to the kinase catalytic domain, they also possess additional domains that are important for enzymatic regulation and substrate recognition. Src families for instance possess additional SH3, SH2 domains. The Abl has an F-actin binding site and a DNA-binding region, Fak possess a ferm domain and a focal adhesion-binding domain which are important for mediating protein-protein interaction [?, ?].

figures/rtk.jpg

(a) Receptor protein tyrosine kinase

figures/nrtk.jpg

**Tyrosine kinase-like group (TKL group)**

The members of this group have close sequence similarity to tyrosine kinases, however, they are mostly serine/threonine kinases and lack the TK-specific motifs. They are mostly diverse with members including receptor and nonreceptor kinases. They comprise of 8 major subfamilies which include IRAK, STKR, RIPK, RAF, LRRK, MLK, MLKL, and LISK.

**STE-group**

The members of this group are classified into three major families. They include STE20 (MAPK4), STE11 (MAPK3) and STE7 (MAP2K). STE stands for "Sterile" and it was originally identified in yeast. The STE kinases sequentially activate each other to then activate the MAPK family.

**RGC-group**

The receptor guanylate cyclase represents the smallest group of the kinases and they consist entirely of pseudo-kinases that lack certain residues that are critical for phosphate transfer [?]. They convert GTP to GMP

**Others**

These include members that lack sufficient sequence similarity to those given above but display unusual phosphorylation properties using ATP and GTP as phosphate donor. Examples include CK2, IKKs.

**Atypical protein kinases (aPKs)**

The atypical kinases represents group of human kinases that lacks similar sequence identity with the ePKs kinase domain HMM profile but have been shown experimentally to have protein kinase activity. Examples includes PIKK family, A6 family, RIO and Pyruvate dehydrogenase kinase [?].



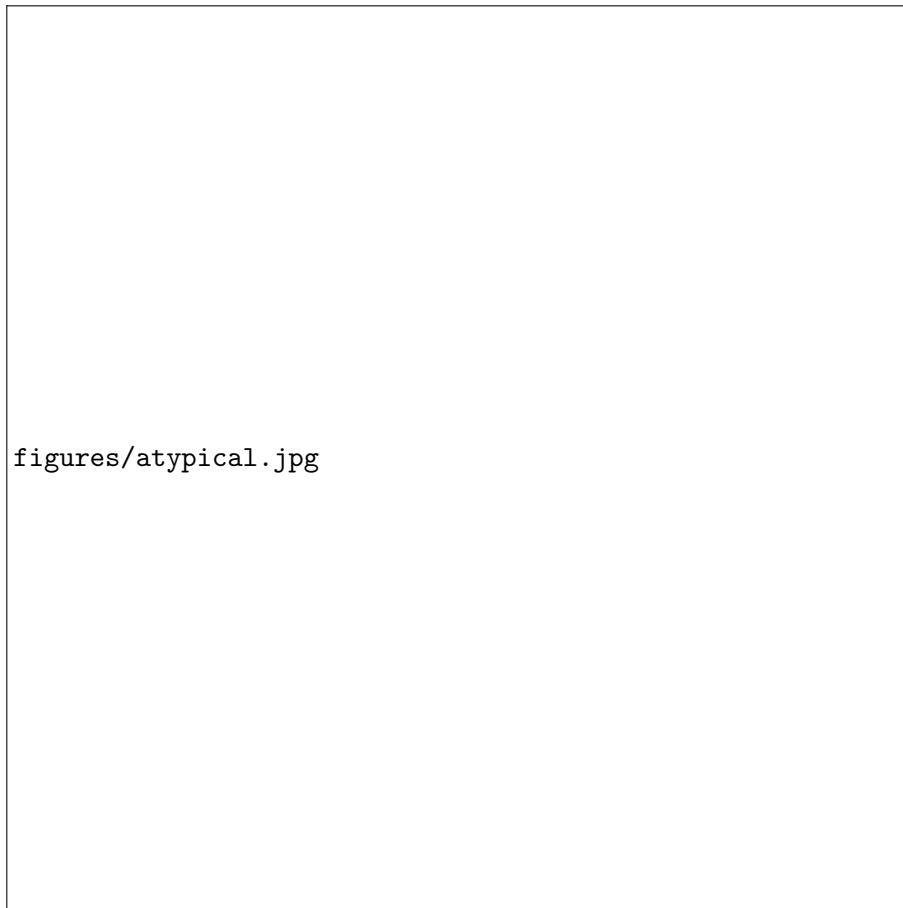


Figure 6: Domain organisation of atypical family of protein kinases. In contrast to classical kinases, the GXGXXG motif of atypical kinases is not involved in MgATP binding but likely involved in peptide interaction.

The crystal structure of the catalytic domain of TRPM7 provides insights into the enzymatic function of the atypical kinases. The comparison of the structure of TRPM7 and PKA catalytic domain reveal some of the major differences between these two. See Figure 7.



Figure 7: Structural comparison of the kinase domains of TRPM7 and PKA. The N-lobe of both PKA and TRPM7 is largely comprised of  $\beta$ -strands while the MgATP binds at the cleft formed from both the N and C-lobes and the binding of Mg in both also involve the conserved P-loop. However, the catalytic loop is not conserved. The GXGXXG motif in TRPM7 contains an extended loop that may play a similar role as the activation loop in classical protein kinase, PKA. Figure obtained from [?].

### Structural Features of Protein Kinase

The kinase domain has two dissimilar lobes joined by a peptide coil called the linker. The N-lobe has about 90 amino acids that are organised into 5  $\beta$ -strands and one helix (C-alpha-helix). This lobe contains the nucleotide binding site that recognises and binds ATP. In the N-lobe, there are highly conserved sequence motifs that are embedded within the first three stands. The first is the GXGXXG motif (Gly-rich loop) which is between  $\beta$ 1 and  $\beta$ 2. This loop folds over the nucleotide and positions the  $\gamma$ -phosphate of the ATP for catalysis. It is the most flexible part of the N-lobe [?]. Another important loop is the P-loop also called the Walker-A motif (GXXGKT/S). Both the glycine rich motif and the P-loop binds to the nucleotide bound phosphate, however, their interaction with the purine is different. For instance, the P-loop does not contact the purine moiety of the ATP while the gly-rich loop connects both  $\beta$  strands that at harbours the adenine ring; the Gly-rich loop is also followed by a conserved Val within the  $\beta$ 2 strand that makes hydrophobic contact

with the base of the ATP [?]. The third important motif is the AxK motif which is found in the  $\beta 3$  strand. The lysine from this motif couples the  $\alpha$  and  $\beta$ -phosphate of the ATP to the C-helix.

figures/kinasedom.jpg

Figure 8: Schematic overview of kinase features. (a). General overview of the organisation of the kinase domain (b) Conserved motifs and residues of the catalytic core of the protein kinases. Taken from [?].

The C-helix serves as a "signaling integration motif" as it connects to different parts of the kinase domain (See Figure 9a). Its C-terminus is connected to the C-lobe by the  $\alpha C$ - $\beta 4$  loop whereas the N-terminus interfaces with the activation loop. The correct positioning of the C-helix is a step required for the activation of the kinase. The distance between the

N-terminus and the activation loop of the C-helix is a measure of the open and closed conformation which is essential for the catalysis [?]. The C-helix is also important in understanding the switch of active to inactive state of protein kinase see section below.

The C-lobe however varies in size, sequence and topology. It is predominately alpha-helical but also contains a few beta strands. It contains the substrate binding groove, activation loop and the catalytic residues. The helical subdomain forms the core of the kinase and the protein/peptide binding surface. The backbone amide of the core helices (D, E, F and H) are well protected from contact with solvent, however, the G-helix is exposed to the solvent. The  $\beta$ -subdomain comprises 4 short  $\beta$  strands (6-9) and contains much of the catalytic machinery for transferring the associated phosphate from the ATP to the protein substrate. The substrate binding site is formed by the hydrophobic residues contributed by the helical core. The activation segment is marked by a conserved DFG (magnesium positioning loop) and APE motif. The activation loop extends from the DFG motif to the Aspartate at the beginning of the F-helix. The length and sequences of the activation segment are the most variable part of the kinase core and this is responsible for turning on and off the kinase [?]. Furthermore, the F-segment extends to the GHI-domain where substrates and regulatory proteins bind. This part is also responsible for stabilizing the active kinase core and also for its allosteric sites (See Figure 9c).

The hinge region of the kinase represents the connecting loop between the N and C-lobe. It contains several conserved residues which provide the catalytic machinery and make up part of the ATP binding pocket. The local spatial pattern (LSP) alignment of protein kinase (a method for comparing two protein structures and identifying spatially conserved residues) revealed two hydrophobic motifs called "spines," that connects the N and C-lobe. These spines give insight into how an active protein kinase is assembled differently from an inactive protein kinase [?]. The R-spine comprises four non-consecutive hydrophobic residues; two from the N-lobe (Leu<sup>106</sup> from  $\beta$ 4 and Leu<sup>91</sup> from C-helix) and the other two from the C-lobe (Phe<sup>185</sup> from the activation loop and Tyr<sup>164</sup> from the catalytic loop). The R spine is therefore an hydrophobic spine that links the two lobes. Using the LSP on the conserved core of the protein kinase, another hydrophobic spine was identified which is called the catalytic(C)-spine. Like the R-spine, it comprises hydrophobic residues belonging to both lobes. In the N-lobe, the Val<sup>57</sup> in the  $\beta$ 2 and Ala<sup>70</sup> from the AxK-motif

as well as the Leu<sup>173</sup> in the C-lobe that docks directly onto the adenine ring of the ATP forming the C-spine. Both spines are anchored to the hydrophobic  $\alpha$ F-helix. Once the R-spine is assembled, and the C-helix is correctly oriented, then the kinase is primed for catalysis. The binding of ATP completes the C-spine and commits the kinase for catalysis [?, ?, ?].

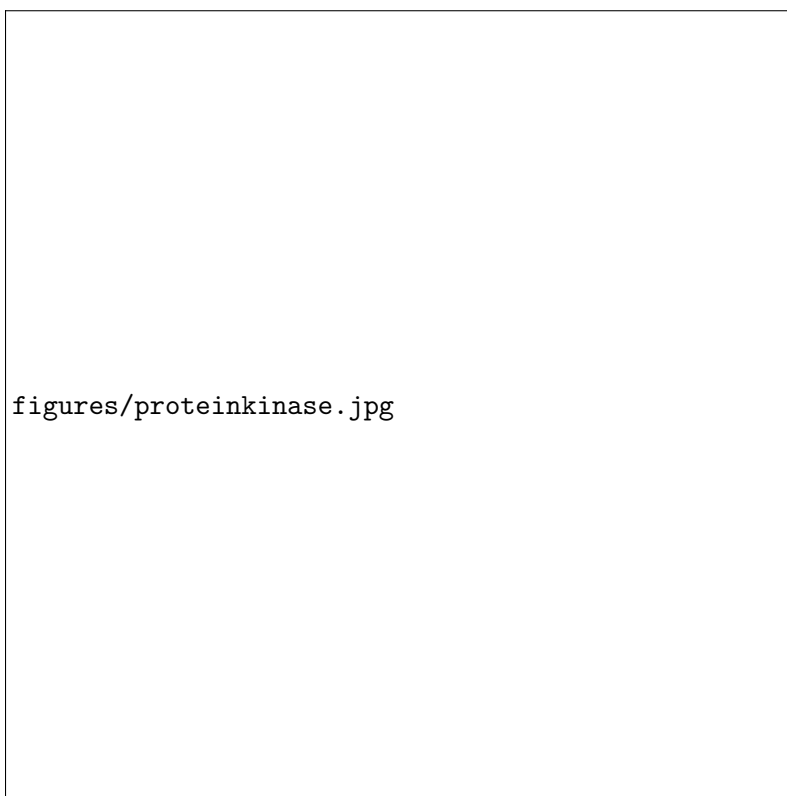


Figure 9: The structure of conserved kinase core. (a) Protein kinase with the characteristic bilobal. (b) In the N-lobe structure, the glycine rich loop coordinates the ATP phosphate binding while the  $\beta$ 3 strand couples the phosphates and the C-helix. (c) Catalytic and regulatory machinery bound to the rigid core of the C-lobe [?]

### Active and Inactive Protein Kinases

The structure of protein kinases reveals the conformational variation of active and inactive kinases. One of the most common forms of inactive protein kinases is the assumption of the DFG-Asp of the activation segment in a out-conformation where the aspartate is directed outward whereas the DFG-Phe is directed inward towards the active site [?]. Structural elements of the kinases show distinct conformation in the active and inactive state. The activation loop for instance is usually in an extended conformation in its active

state whereas it is disordered with the loop collapsed to block the substrate binding, in the inactive state (see Figure10). The phosphorylation of the residues within the activation loop activates the kinases [?].



Figure 10: The active and inactive conformation of LCK and Src respectively. (a) Active conformation with activation loop adopting an extended conformation while it folds in the inactive c-Src kinase domain (b). Figure taken from [?].

Furthermore, the presence of a salt bridge between the  $\beta 3$ -lysine and the  $\alpha C$ -glutamate, together with the formation of the R- and C-spine, are the hallmarks of an active kinase domain while inactivation involves the disassembly of the R-spine. The rotation or shift of movement of the  $\alpha C$ -helix also cause a switch from an inactive to an active kinase as the  $\alpha C$  adopts an in-conformation in its active state and an out-conformation in its inactive state. [?, ?].

## Kinase Inhibitors

The kinases are quite diverse in their primary sequences. However, they share a great degree of similarity in their 3D structure most especially in their catalytic site where the ATP-binding cavity is found; a  $\beta$  sheet containing N-lobe as well as  $\alpha$  helix dominated C-terminal (C-lobe) with a connecting hinge region. ATP binds in the cleft between the N and C lobes and therefore most kinase inhibitors interact with this region to perturb the binding of ATP [?]. There are several kinds of inhibitors that are being exploited in targeting protein kinases. These inhibitors differ in their mode of binding and the mechanism of action exhibited upon binding. The kinase inhibitors can either bind covalently or reversibly.

The nonreversible (covalent) inhibitors bind irreversibly with the reactive nucleophilic cysteine or lysine residue close to the ATP-binding site resulting in the blockage of ATP binding and leading to irreversible inhibition. Example of such a drug in clinical trial is the AVL-292 which is a tyrosine kinase inhibitor which covalently binds to the Bruton tyrosine kinase (BTK)[?], ibrutinib targets BTK as well, while afatinib targets the gefitinib resistant EGFR [?].



Figure 11: Afatinib co-crystal structure with wild-type EGFR (PDB ID: 4G5J) and mutant T790M EGFR (PDB ID:4G5P). Afatinib binds to the kinase domain in its active conformation and forms a hydrogen bond with the backbone NH of Met793 and also forms covalent interaction with the sulphur of Cys797. Figure obtained from [?]

The reversible (non-covalent) inhibitors on the other hand can be classified into several types, based on their interaction with binding pocket and the DFG motif (hinge region). The type-1 inhibitors are ATP-competitors that bind to the active form of the enzyme with the aspartate residue of the DFG motif facing the active site of the kinase (DFG-in conformation). The conserved Phe of the DFG-motif is buried within the hydrophobic pocket of the groove between the N and C-lobes. Most of the compounds that target this active conformation have been selected using enzymatic assays that select ATP mimetics with the highest inhibitory activity for the kinase [?]. Classical examples of such approved inhibitors include gefitinib, dasatinib, erlotinib and sunitinib.



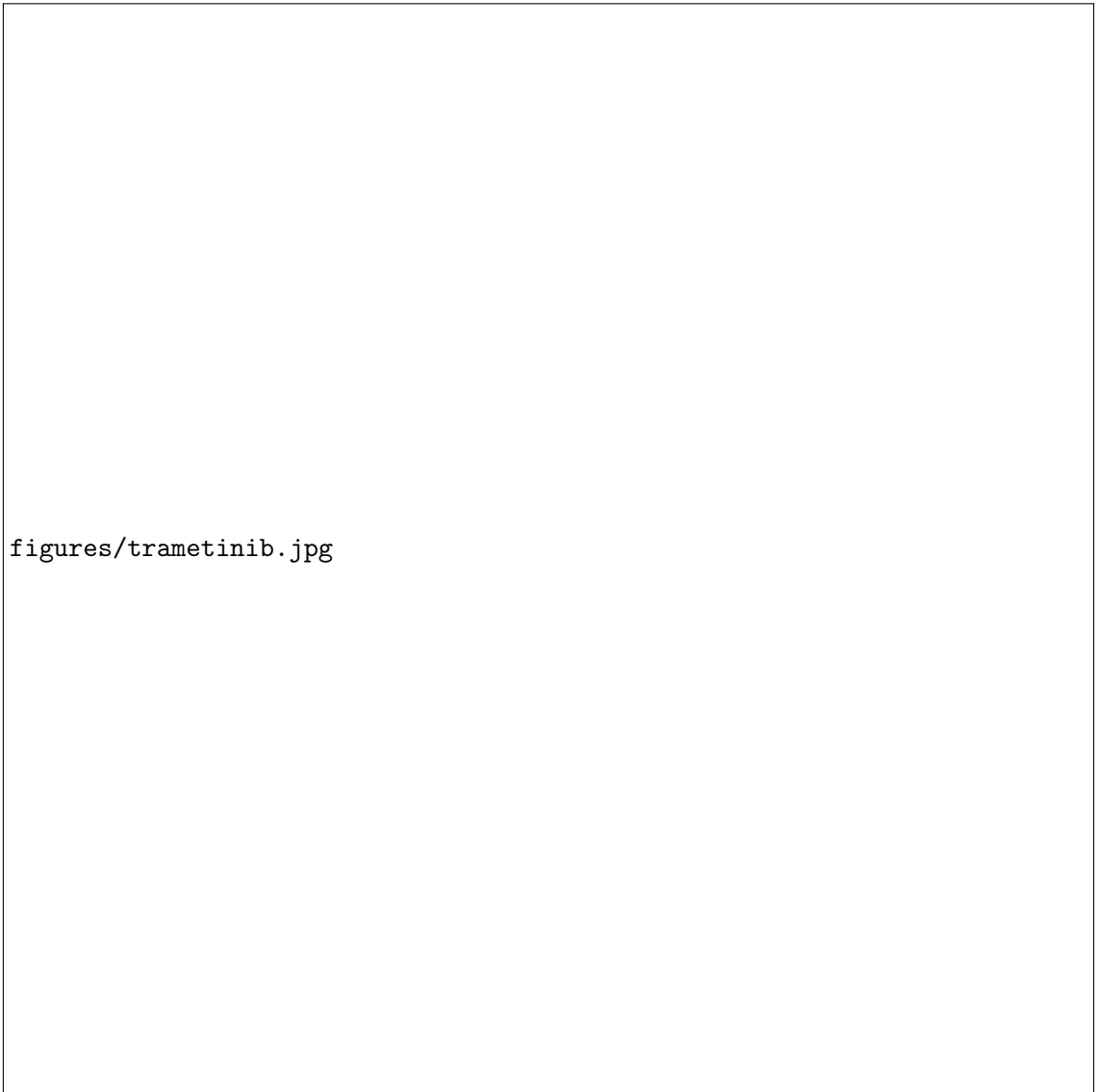


Figure 12: Crystal structure of EGFR tyrosine kinase domain (TKD) bound with inhibitors. (A) Erlotinib bound with EGFR-TKD in the inactive state (B) Lapatinib with inactive EGFR-TKD (1XKK) (C) Erlotinib with the active EGFR-TKD (1M17). Figure obtained from [?]

The type-2 inhibitor binds to the inactive form of the enzyme with the aspartate residue of the DFG motif protruding outward from the ATP-binding site of the kinase. The transition from the DFG-in to DFG-out conformation exposes the hydrophobic pocket adjacent to the ATP-binding site and this is utilized by the type-2 inhibitor to lock the kinase in an inactive conformation [?]. The type-2 are generally less promiscuous as compared to the type-1 inhibitors. Examples of FDA-approved type II kinase inhibitors include imatinib, nilotinib, and sorafenib [?]. The type-1 and 2 inhibitors however face competition with the millimolar concentration of ATP *in vivo* as well as lack of selectivity due to the extensive adenosine binding cleft [?]. There have therefore been efforts directed towards kinase inhibitors with high selectivity, high affinity and less side effects from usage.

The type-3 inhibitors are a heterogeneous group of kinase inhibitors that bind to allosteric or remote sites on the kinase. These inhibitors mostly do not bind at the ATP-binding

sites and have no physical contact with the hinge and they have been shown to exhibit the highest form of selectivity by exploiting the binding and regulatory sites that are specific to a particular kinase [?]. The combinations of the structural elements in the kinases such as the C-helix's DFG-in and out state, A-loop, G-loop, C-terminal elements as well as regulatory domains can be exploited to design selective inhibitors with clear advantage over the type-1 and 2 inhibitors [?]. Examples of approved type-3 inhibitors include cobimetinib, trametinib, selumetinib, binimetinib and rapamycin. Type-3 inhibitor of MEK1 binds to the adjacent pocket to the ATP-site which is referred to as the "allosteric back pocket-DFG-in" in the presence of ATP and "allosteric back pocket-DFG-out" in the absence of ATP [?].



figures/trametinib.jpg

Figure 13: MEK kinase inhibitor binding mode (A). The chemical structure of trametinib (B). The binding mode of trametinib with MEK1 (C). Tak-733 co-crystallized with MEK1 (PDB ID: 3PPI) ATP is shown in cyan and Tak-733 in magenta. Figure taken from [?]

The type-4 allosteric inhibitors bind at allosteric sites that is distant from the the ATP-binding site. A unique example is the AktI-1/2 targeted inhibitor that inhibits Akt isoforms 1 and 2 kinases. Interestingly, these inhibitors has no inhibition against PH-domain mutants which suggest that the PH domain is required to exert their activity. This clearly shows that the inhibitor interacts with both the catalytic domain and the PH domain and prevents the activation of the upstream kinase PDK1 [?]. Other types of allosteric protein kinase inhibitors includes the type-5 which are also referred to as bivalent or bi-substrate inhibitor. The bivalent inhibitors tend to have high affinity and

more selectivity for targeted therapy. The design of such inhibitors involve the use of an appropriate linker to couple allosteric site inhibitor with kinase active site binding agent to achieve improved selectivity from the non-ATP directed inhibitor [?]. Other examples of kinase inhibitors is the hybrid-type having both type I and II features. The field of allosteric kinase inhibition is a rapidly evolving field with the FDA-approval of trametinib as well as several other allosteric inhibitors that are in clinical trials [?].



Figure 14: Kinase structure and the various types of inhibitor schematics. (A). Co-crystal structure of PDK1 with ATP (adenine and ribose in green, phosphate in orange) Enlarged area shows hydrogen bond in red, hinge and hinge residues in green backbone, P-loop and P-loop residues in brown-orange, Asp residue of the DFG motif and the activation loop in grey [PDB ID:4RRV, 1.41Å], (B) The four types of reversible binding mode. Figure taken from [?]

Allosteric inhibition offer some advantages such as high selectivity and ability to overcome drug resistance as most drug resistance to small molecule kinase inhibitors occurs frequently around the hinge region. However, some debated opinion about allosteric

inhibitors include; mutation-related resistance may occur at the allosteric sites as they are not as essential for kinase function as the ATP binding sites. Also, as a result of the hydrophobic properties of most allosteric pockets, the allosteric inhibitors are lipophilic compounds and this may result in poor bioavailability, and poor solubility. Another major challenge is the limited numbers of structures for allosteric-inhibitor-bound kinases to help in the comparison of the induced changes associated with the on/off-bound state of the enzymes. This may be due to the fact that these sites are involved in protein-protein and protein-peptide interactions and the transient nature of such interactions creates difficulty in solving the structures. [?].

### **Understanding the promiscuity of Protein Kinase Inhibitors**

Protein kinase inhibitors are generally considered promiscuous because of their lack of specificity and ability to interact with several kinases and families, due to the presence of the conserved ATP-binding site where the kinase inhibitors interact with. Promiscuity is defined as the ability of a compound to specifically interact with more than one target (the target of interest for which it was designed [?]. Hu et al classified the protein kinase inhibitors into single and multiple kinase inhibitors based on the numbers of targets the PKI compounds in ChEMBL database were active against. Furthermore, they also assessed the promiscuity of a kinase for several structurally diverse compounds and found that many kinases recognise structurally diversified compounds [?].

In the work earlier reported by [?], they understood this concept of cross reactivity and promiscuity associated with kinase inhibitors. However, they identify some "selective filters" can may be used to distinguish safer drugs and targets from the list of paralog targets the drugs might be associated with. This selective filters are called dehydrons which are also referred to as "wrapping patterns" of the protein targets. They are the intermolecular protection of backbone hydrogen bonds from the hydration of amide and carbonyl groups. The wrapping defects could be informed not only from the structure of the protein kinase but also the sequence. The incorporation of dehydrons into drugs designs have been shown to reduce the numbers of cross-reactivity and improve the specificity of protein kinase inhibitors [?].

## Objectives

We have previously demonstrated the importance of using CATH-FunFams as a reasonable annotation level for drug-domain interaction and also used network analysis to associate the propensity of side effects with the network properties of these families through our druggable FunFam approach [?]. In this study, we are focusing on protein kinase inhibitors, a set of molecules that inhibit the activities of kinases. Using a set of publicly available protein kinase inhibitors, we intend to associate the targeted kinases with our domain-families (FunFams) and subsequently measure some network characteristics to distinguish FunFams with side effects from others. We will also be exploring some similarity measures and structural coherence to shed more lights on possible repurposing of protein kinase inhibitors to members of a given FunFam.

## Results and Discussion

### GSK-Protein Kinase Inhibitor

The Published Kinase Inhibitor Set (PKIS) is a collection of 367 compounds that have been made available by GSK to the external community [?, ?]. These compounds have been annotated with protein kinase activity [?]. These compounds are of various chemotypes and they are openly available from the ChEMBL database (ChEMBL release 23) [?]. The PKIS are active against some known target kinases and can be extended to other new target kinases. We collected only a subset of the PKIS that indicate an inhibitory activity level above 50% as [?, ?] had reported this threshold as appropriate for considering the inhibition of kinase catalytic activity. This gave a unique set of 205 protein kinase inhibitors that were distributed across 133 protein kinases. This covers about 60% of the entire PKIS set and thus a reasonable dataset to consider for a network assessment and characterization of protein kinase inhibition.

### Target Promiscuity of Protein Kinase Inhibitors

The PKIS dataset obtained from ChEMBL database was analysed to study the association of kinases (targets) with inhibitors in order to understand the inherent promiscuity associated

with the protein kinase inhibitors set.

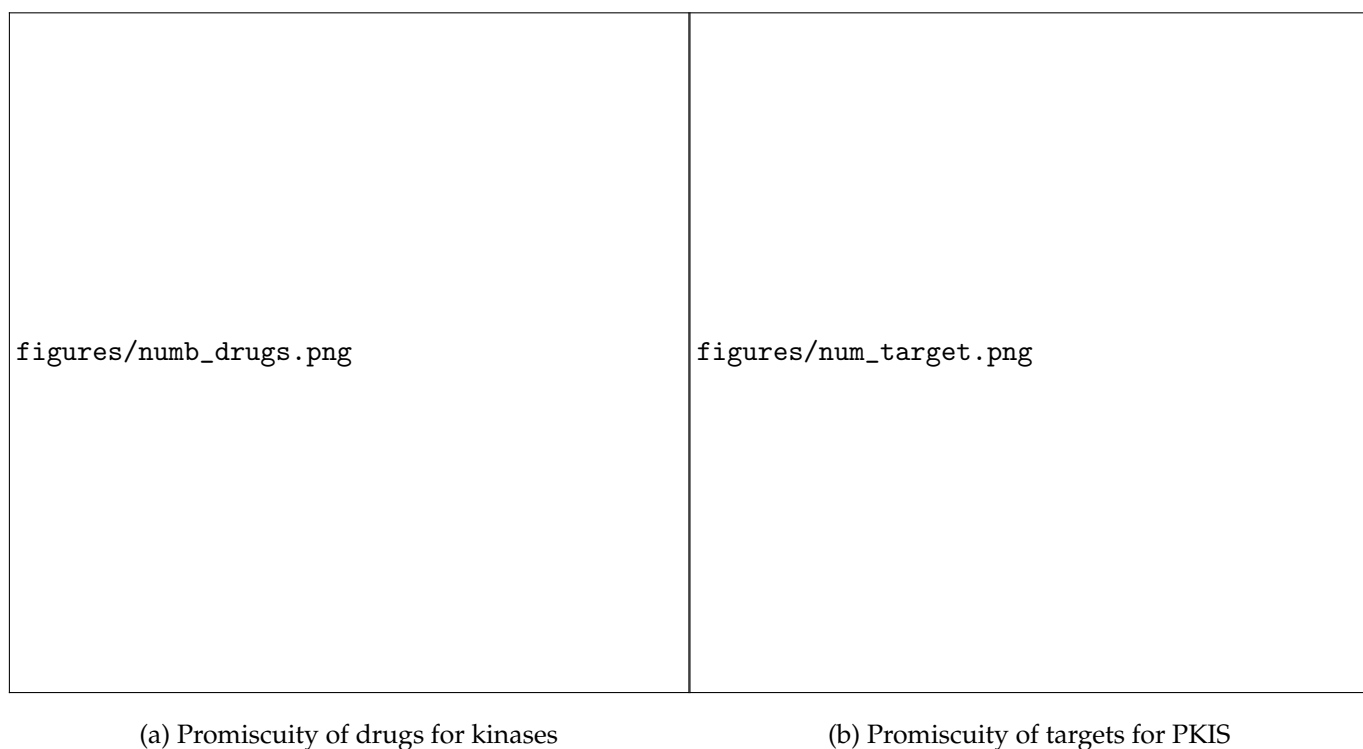


Figure 15: Distribution of kinase-drug interaction and drug-kinase interaction for the GSK-PKIS sets

Figure 15 shows the promiscuity of the kinases for inhibitors as well as inhibitors for kinases. Protein kinase inhibitors interact with more than one kinase while some kinases could also interact with more than one protein kinase inhibitor, which indicates the non-specificity of these kinase sets.

### Mapping kinases to Pfam-FunFam

Protein kinases have been divided into 9 groups by [?] as explained in the introductory section. We associated all human kinases with our functional families (groups of evolutionarily related, structurally and functionally coherent protein families). We scanned all human kinase sequences from Pfam against Pfam-FunFams library using HMMer3. Pfam-FunFam data from the Gene3D database was used as Pfam provides sequences that cover the entire kinase catalytic region, whilst CATH divides the kinases into the N and C lobe domains.

1277 human protein kinases sequences obtained from Pfam was clustered at 90% sequence

identity to obtain a unique non-redundant human kinase set using CD-HIT clustering algorithm. The 741 unique kinases obtained were distributed amongst 130 Pfam-FunFams. The identified Pfam-FunFam represents 46% of the entire Pfam-clan as the human protein kinases were associated with 16 of the 35 Pfam clans. We mapped the Pfam-FunFam to the human kinome tree and found that the our domain-family classification are well annotated to the various groups with no overlap between the already identified grouping by [?] as shown in figure 16 below.

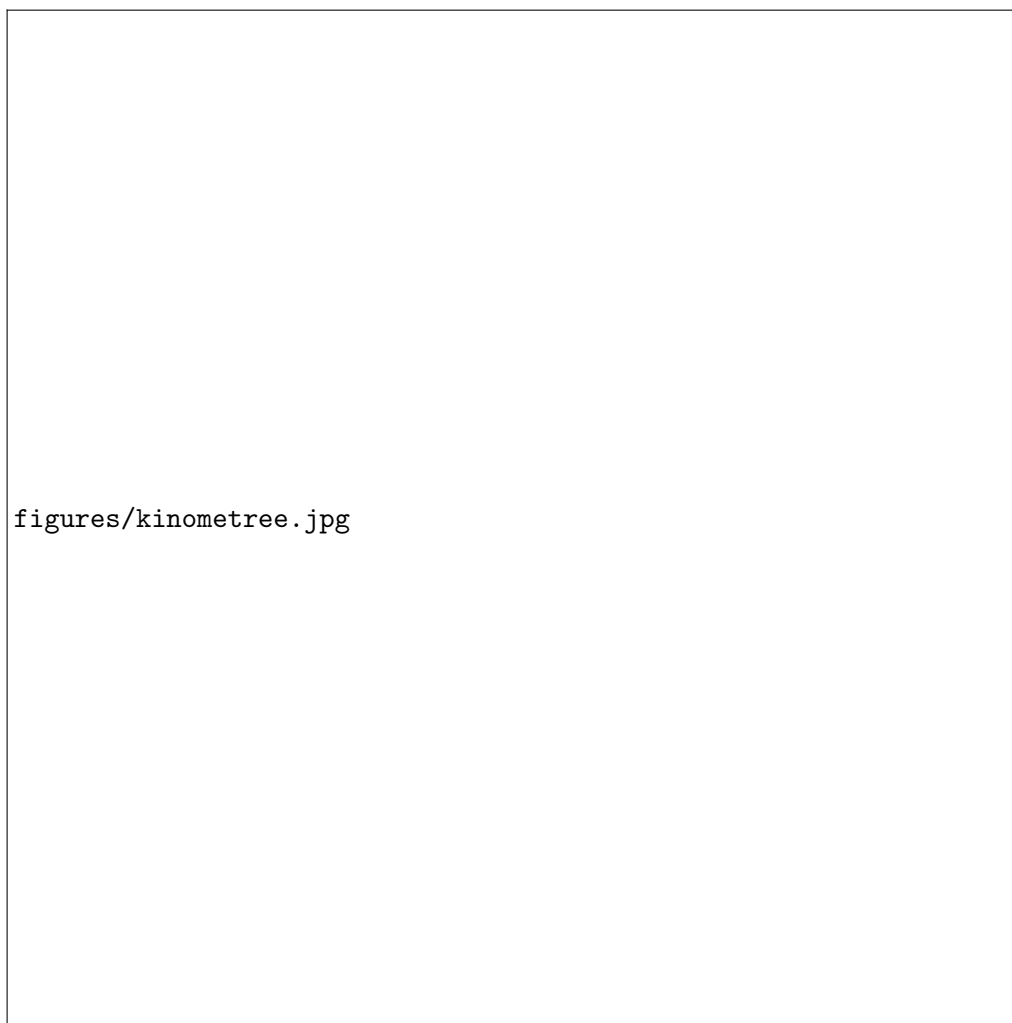


Figure 16: Pfam-FunFam distribution across the human kinome. The kinome tree was adapted from [?]. The mapping of Pfam-FunFam to the group "Others" is not shown in this figure

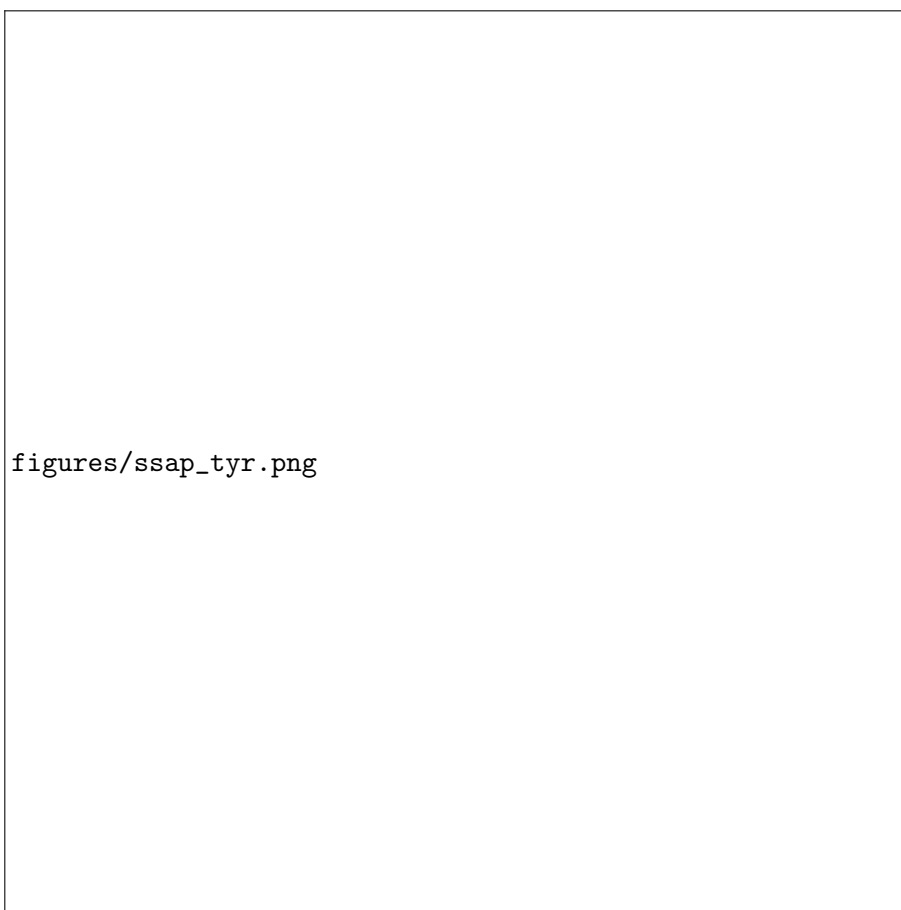
We then mapped the drug-targets of the PKIS obtained from ChEMBL to the Pfam-FunFams at 50% inhibition level with an affinity of  $0.1\mu\text{M}$  and found that they are associated with 37 of the 130 Pfam-FunFams which represents about 30% of the human kinase



Pfam-FunFam obtained. We speculated that this 30% represents the kinases most targeted by the pharmaceutical industry based on their relevance to human disease and those most studied as the current research in kinase therapeutics indicate that only 10-15% is being targeted, we also observed this ratio from the entire GSK-PKIS set being used in this study.

### **Conservation of human kinase relatives in the Pfam-FunFams**

We measured how conserved the kinases across the Pfam-FunFam are. The human relatives of the Pfam-FunFam were mapped to structure by SIFT-ing through UniProt sequences to PDB while the domain region were specified using Pfam. The structures were therefore evaluated for structural conservation by measuring RMSD using the SSAP algorithm. The groups were divided into the tyrosine-kinase Pfam-FunFam and serine/threonine Pfam-FunFams. The distribution of the RMSD across the two Pfam-FunFam groups is shown in figure 17 below.



(a) RMSD across the Tyr Pfam-FunFam



(b) RMSD across the Ser/Thr Pfam-FunFam

The figure 17 shows the conservation of the relatives in these families as the observed RMSD was relatively below 2 in both the tyrosine kinases as well as the serine/threonine kinase FunFams. This therefore agrees with the overall view of the structural conservation of the kinases. However, some families show that Pfam-FunFams with members having high RMSD score which tend to distort the overall RMSD measure and indicating lack of structural coherence with relatives of the FunFams.

A deeper insight into one of such families was carried out using the Pfam-FunFam PF07714.FF13122 which gave an overall RMSD below 2Å but contains outliers that influence some members having RMSD as high as 5Å in comparison to other members with low RMSD. The reason for this could be attributed to the multidomain architecture of the relatives of the Pfam-FunFam as majority of the relative in this protein have additional SH2-domain while others either have an immunoglobulin-domains or no additional domain.

### **Enrichment test of Pfam-FunFams associated with drug targets**

[?] reported the thorough characterization of the Published Kinase Inhibitor Sets (PKIS) and identified them as the chemical starting point for probing orphan kinases. They illustrated the utility of these compounds for developing selective inhibitors against untargeted kinases LOK and SLK. Thus, the use of domain families could help increase the coverage of potential targets of the kinase-inhibitor set as the targeted kinases are about 10-15% . We used our enriched FunFam program (see chapter two) to test for the overrepresentation of drugs in our FunFams. This involves using a binomial test followed by multiple testing for correction to determine the most appropriate FunFam with which a Kinase inhibitor associates. We found that PKIS-drugtargets were overrepresented in 30 Pfam-FunFam at  $p\text{-value} \leq 0.05$ . 109 PKIS were found to be associated with the 30 Pfam-FunFams.



Figure 18: Distribution of drugs in the overrepresented FunFams

As shown in figure 18, we observed that 70% of the enriched Pfam-FunFams were associated with more than 2 kinase inhibitors from the PKIS set. This enriched Pfam-FunFam dataset was used for further analysis of the protein-kinase inhibitor targets on protein-protein interaction network.

The Pfam-FunFam-drug interaction was represented in network for visualization in Cytoscape as shown in figure 19 while the names of the overrepresented Pfam-FunFams were identified using our in-house program for the UniProt description of protein sequences as shown in table 1.

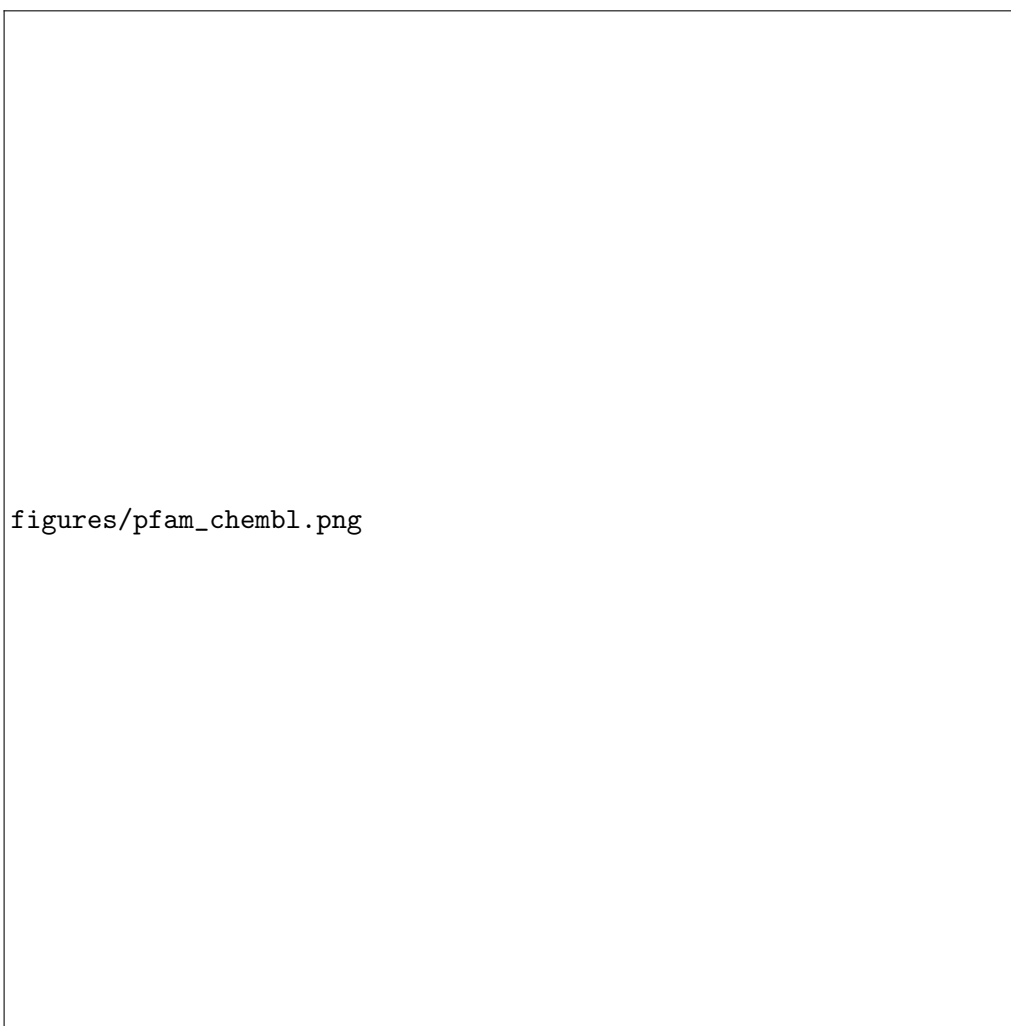


Figure 19: Pfam-FunFam-drug interaction network. In this network, the green coloured square node represents the CheMBL protein kinase inhibitors while the purple coloured circle nodes are the Pfam-FunFam whose relative are quite coherent with a mean  $\text{RMSD} \leq 2$  while the blue coloured circle are Pfam-FunFam with  $\text{RMSD}$  value  $\geq 2.5$ . The size of each node reflects the numbers of targets (relatives) in each family. Also labelled are some interesting families with relatives  $\geq 5$  and interacting with at least 5 drugs.

Table 1: The overrepresented FunFams and the number of drugs associated with each FunFam.

Pfam-FunFam	Uniprot-Description	No of Drugs
PF00069.FF13563	Ribosomal protein S6 kinase alpha-5	3
PF00069.FF25168	Protein kinase B beta	4
PF00069.FF31429	MAP/microtubule affinity-regulating kinase 3	1
PF00069.FF31839	Serine/threonine-protein kinase 4	2
PF00069.FF37132	Serine/threonine-protein kinase BRSK2	1
PF00069.FF62179	Serine/threonine-protein kinase pim-1	1
PF00069.FF62314	Mitogen-activated protein (MAP) kinase	4
PF00069.FF62318	p90 ribosomal S6 kinase	6
PF00069.FF62341	Cell division protein kinase	4
PF00069.FF62345	Calcium/calmodulin-dependent serine/threonine-protein kinase 1	5
PF00069.FF62348	Microtubule-associated serine/threonine-protein kinase 2	10
PF00069.FF62351	CBL-interacting serine/threonine-protein kinase 1	1
PF00069.FF62355	Non-specific serine/threonine protein kinase	1
PF00069.FF62363	Dual specificity tyrosine-phosphorylation-regulated kinase	10
PF00069.FF62561	Cyclin-dependent kinase 6	1
PF00069.FF62583	Testis-specific serine/threonine-protein kinase 1	2
PF00069.FF62585	Serine/threonine-protein kinase Nek8, related	3
PF00069.FF62682	MAP kinase-interacting serine/threonine-protein kinase	1
PF00069.FF62684	Glycogen synthase kinase 3	17
PF00069.FF62706	Ttk family protein kinase	2
PF00069.FF62744	Casein kinase I alpha	3
PF00069.FF62866	Leucine-rich repeat serine/threonine-protein kinase	2
PF00069.FF7038	Cyclin-dependent kinase 2	8
PF07714.FF123	Platelet-derived growth factor receptor alpha	23
PF07714.FF13026	Receptor protein-tyrosine kinase	17
PF07714.FF13065	Receptor protein-tyrosine kinase	4
PF07714.FF13096	Tyrosine-protein kinase	1
PF07714.FF13122	Non-specific protein-tyrosine kinase	1
PF07714.FF13154	LIM domain kinase 2	2

## Network Analysis of the Pfam-FunFams Enriched with Drug Targets

The representation of proteins on a network gives a view of the information flow and interactors for biological process. We obtained a functional protein association network for human from the STRING database version 10.0. We filtered this data and only used those edges with a combined and database score of  $\geq 800$  which correspond to those PPI with high reliability. This gave 219,608 physical interactions between 10,430 proteins. We extracted the largest connected subgraph and then computed the node centralities

for drug targets from the enriched Pfam-FunFams using centrality measures such as the betweenness centrality and PageRank.

### Centrality Measure

Centrality measures identify important nodes relative to other nodes within the network. Such measures include the degree, betweenness centrality as well as the PageRank. The degree of a node is the number of connections (edges) it shares with other nodes. The betweenness centrality (BC) is however the fraction of the number of shortest paths that pass through each node. The BC measures how often a node occurs on all the shortest paths between two nodes. Therefore, a node with high BC influences the flow of information in the network.

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where  $V$  is the set of nodes,  $\sigma(s,t)$  is the number of shortest  $(s,t)$ -paths, and  $\sigma(s,t|v)$  is the number of those paths passing through some node  $v$  other than  $s, t$ . If  $s = t$ ,  $\sigma(s,t) = 1$ , and if  $v \in s, t$ ,  $\sigma(s,t|v) = 0$

We measured the topological properties of the targeted kinases and compared them to all human kinase and all proteins in a given human functional network. we defined here "Hubs" as the top 20% nodes with highest degree (connections) while "High-BC" are those top 20% nodes with the highest betweenness centrality. The "Bottlenecks" we have defined as those nodes within the "High-BC" but excluded from "Hubs" i.e. those with low degree connectivity. The "Hubs & High-BC" are those set of nodes in the Hubs group with high betweenness centrality measure.

Table 2: Topological analysis of kinases in a functional protein network

Groups	Hubs (%)	Bottlenecks (%)	High-BC (%)	Hubs & High-BC (%)
Targeted Kinases	22.64	2.8	21.69	24.52
All-human kinases	8.08	2.15	7.14	9.3
All-human proteins	19.68	11.56	8.12	19.68

The measure of degree and betweenness centrality are some of the most profound topological properties of nodes in a protein interaction network. The bottlenecks in a protein

functional network represents key connectors and surprisingly show functional and dynamic properties [?]. The table 2 shows the comparison of the targeted kinases, all human kinases and all human proteins in a functional network. The proportion is a measured relative to each group as;

$$\frac{\text{Number of Hubs or bottlenecks in targeted set} \times 100}{\text{Total number proteins in targeted set}}$$

The result therefore shows that only a small percentage of the targeted kinases are bottlenecks. Our observation shows that on average, the betweenness centrality is lower in the targeted kinases compared to all the proteins in the network, however, the violin oplot in figure 20 indicates that some of the targeted kinase have a higher betweenness centrality measure. We hypothesize that bottlenecks will be a good drug target as they are central in the network but associated with less nodes (thus less functional disintegration expected). However, there is still a lot of debate on this opinion as some studies have suggested bottlenecks to be associated with side effects.





Figure 20: Violin plot showing bottlenecks of all proteins in a functional protein network compared to the kinases in the network. We observed a significant difference between the bottlenecks of all proteins compared to the kinases (pvalue =  $1.394e-14$ ) but no significance observed in the bottlenecks associated with the targeted kinases and all human kinases (pvalue=0.740).

Furthermore, we consider the Pfam-FF which we have identified using our overrepresentation analysis and measure the topological properties of the relatives, hence, calculating the proportions of bottlenecks in each domain-family as a fraction of the entire relatives in the given funfam.

Table 3: The topological analysis of the relatives within a given Pfam-FF

<i>Pfam-FF</i>	<i>Bottlenecks (%)</i>	<i>Hubs (%)</i>
PF00069.FF62561	0	0
PF00069.FF62351	3.77	7.55
PF00069.FF62318	1.2	16.87
PF00069.FF62355	0	3.49
PF00069.FF62314	2	16
PF07714.FF212	0	37.5
PF07714.FF13154	0	9.52
PF07714.FF13122	0	21.43
PF00069.FF62866	100	0
PF07714.FF13026	0.51	8.08
PF07714.FF13065	0	1.92
PF00069.FF37132	0	0
PF00069.FF62706	0	0
PF00069.FF62744	7.69	0
PF00069.FF25168	0	50
PF00069.FF62682	0	0
PF00069.FF62684	0	16.67
PF00069.FF62341	0	5.13
PF00069.FF62363	8	0
PF00069.FF62345	4.84	0
PF00069.FF31429	6.25	0
PF00069.FF62348	0	3.57
PF07714.FF123	0	44.44
PF00069.FF7038	0	14.29
PF00069.FF13563	0	0
PF00069.FF62583	0	6.06
PF00069.FF62585	0	2.94
PF07714.FF13096	0	26.67
PF00069.FF31839	0	0
PF00069.FF62179	0	0

The result obtained in table 3 indicate that most of the relatives of the kinase-domain Pfam-FunFam are majorly within the hub-class. This indication is not far-fetched as it suggest the high connectivity of the kinase and interaction with several substrate.

### Kernel Similarity of Drug Targets in Pfam-FunFams in a Protein Functional Network

Following our initial hypothesis that FunFams whose relatives aggregate in a network neighborhood are likely to be enriched with potential targets and free of off-targets, we assessed the kernel similarity of all the Pfam-FunFam kinases. The kernel similarities have a score ranging from 0-1 with 1 indicating a high similarity and 0 showing no similarity.

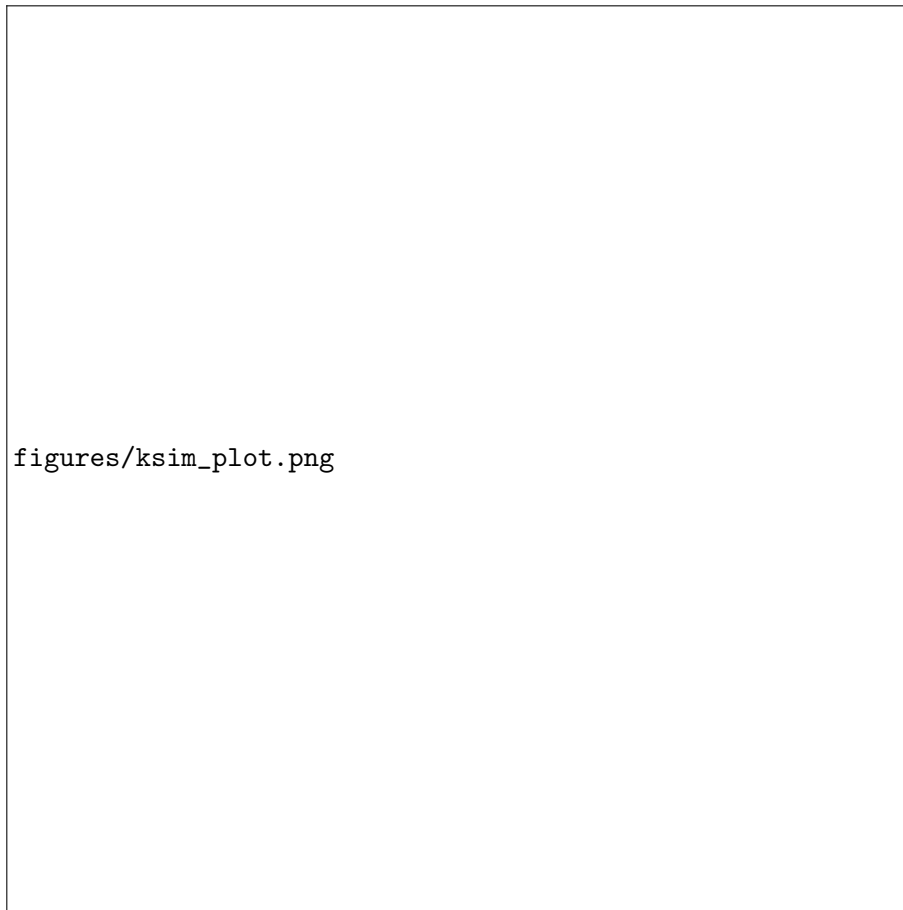
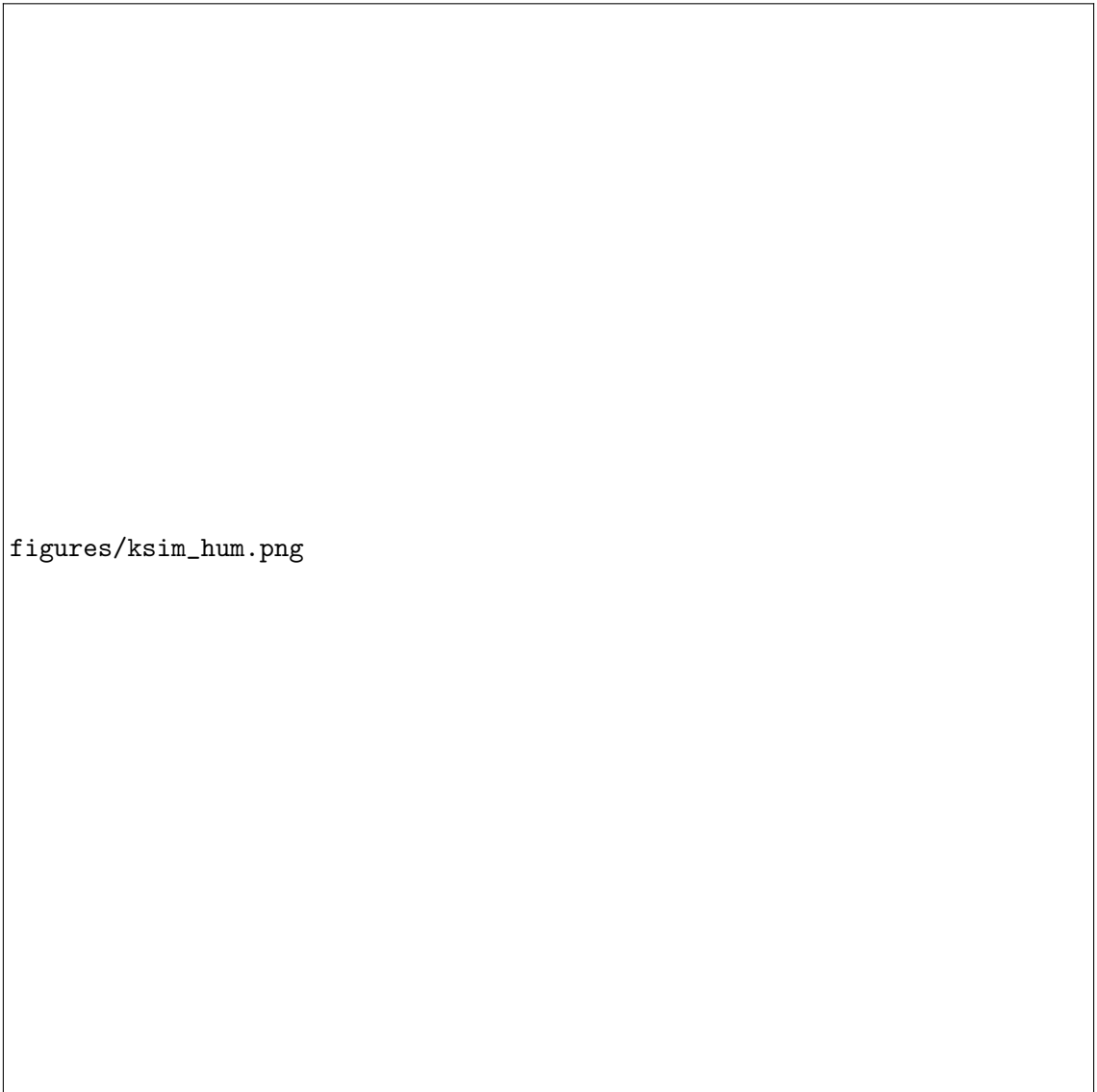


Figure 21: Distribution of the kernel similarity measure across the Pfam-FunFam.

The results indicate that most of the Pfam-FunFam relatives are scattered across the functional protein network with majority of the Pfam-FunFam family having a similarity score lower than 0.5. This may explain one of the reason for the side effects associated with protein kinase inhibitors as their targets arequite diverse and could affect many biological pathways i.e. elicits multiple pharmacological responses in a given organism.

We then compared the kernel similarity of targeted kinase, all human kinases and the all the human proteins in the protein functional network.



figures/ksim\_hum.png

Figure 22: Cumulative probability plot of the kernel similarity of all human targets compared against PKI-targets

We observed as shown in figure 22 that the targeted kinases has a lower kernel similarity as compared to the distribution of the kernel similarity of the all human proteins in the network however, there is no difference between the targeted kinases and all the human kinases as they show the same distribution as observed in the figure 22. To furthermore corroborate our observation, we used another modular measure called DS\_score which was reported by [?]. The DS\_score measures the network distance of proteins belonging to a group of drug-targets and compares their closeness to random expectation. We observed that there is a significant difference between DS\_score of the target proteins compared to random ( $P \leq 1.116 \times 10^{-8}$ , *Mann – Whitney* U-test). The lower the DS\_score, the more aggregated the proteins in a network. This observation hence shows that the targeted

kinases are quite clustered in a protein functional network compared to random proteins.

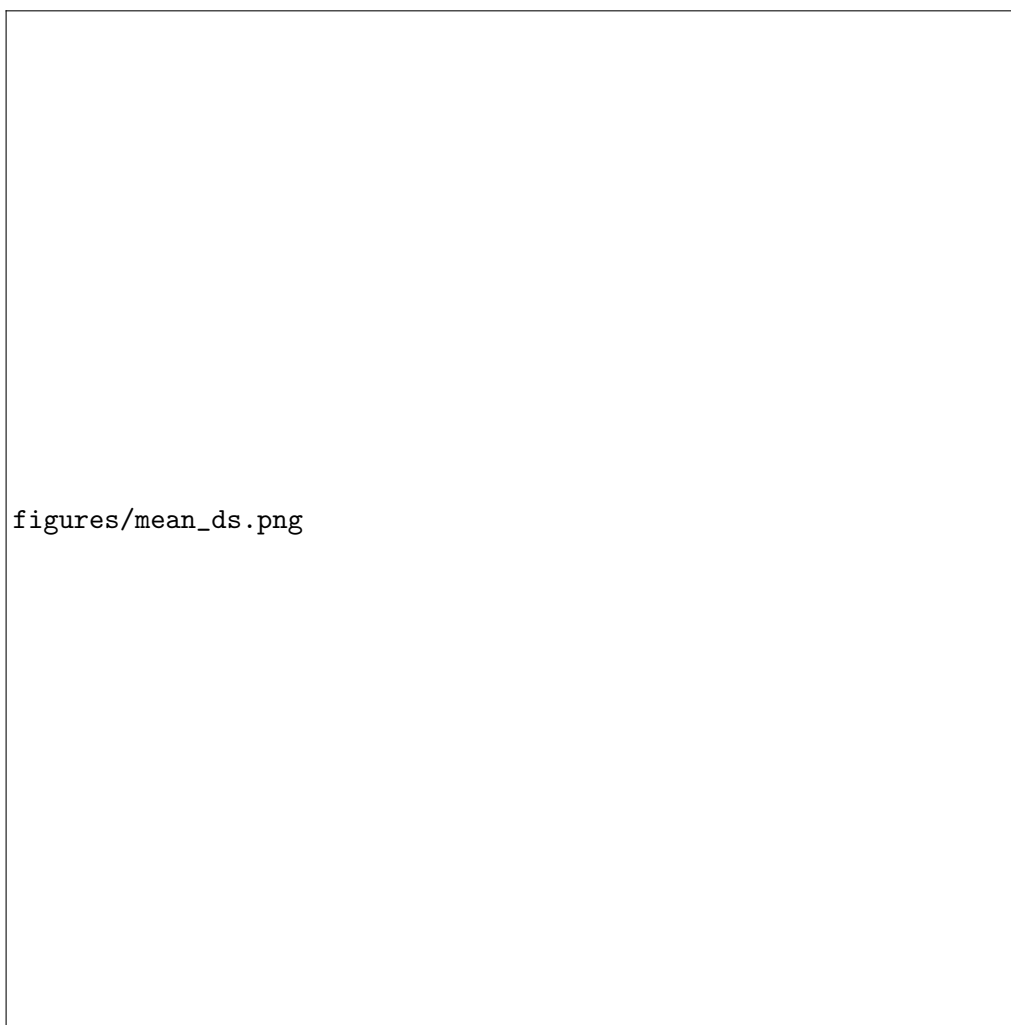


Figure 23: Distribution plot of the DS-measure of PKI-targets compared against random proteins in a protein functional network

### **Structural coherence of the binding site of the enriched Pfam-FunFam relatives**

The work published by [?] reported some of the solved structures of the binding of inhibitors from the PKIS with some kinases. The crystal structures of the inhibitor (ChEMBL237571) with the lymphocyte-oriented kinase (LOK) in the inactive DFG-out state (PDB-ID:4USD) and active DFG-in state (PDB-ID: 4USE) have been deposited in PDB. Therefore, using this example, we were interested in observing the binding site conservation across members of the Pfam-FunFam this target belongs to.

The target for this inhibitor belongs to the Pfam-FunFam PF00069.62355 (a STE-group kinase) that has about 80 relatives. Using the SIFTS mapping of UniProt-sequences to

PDB structures, we were able to identify 15 members (about 18%) of this family with PDB structures. In case of multiple structures of the same kinase, firstly, we find the structure without ligands, otherwise we chose the best resolved structure for the particular kinase.

Table 4: The list of Pfam-FunFam (PF00069.62355) relatives with structural information

UniProt-ID	PDB-ID	Resolution (Å)
O95819	4u3y	1.45
Q9P289	3ggf	2.35
Q9P286	2f57	1.80
O96013	2j0i	1.60
Q8IVH8	5j5t	2.85
Q9Y6E0	3a7f	1.55
Q13153	1yhv	1.80
Q9UKE5	2x7f	2.80
Q9H2G2	2j51	2.10
Q9NQU5	2c30	1.60
Q99759	2o2v	1.83
Q9Y2U5	5ex0	2.70
Q13177	3pcs	2.86
O00506	2xik	1.97
O94804	2j7t	2.00

We structurally superposed the members of this family to give a pictorial representation of the conservation of the relatives and binding residues interacting with the CHEMBL237571. This family was found to be structurally coherent as the SSAP score gave an RMSD of  $1.11 \pm 0.48$ .

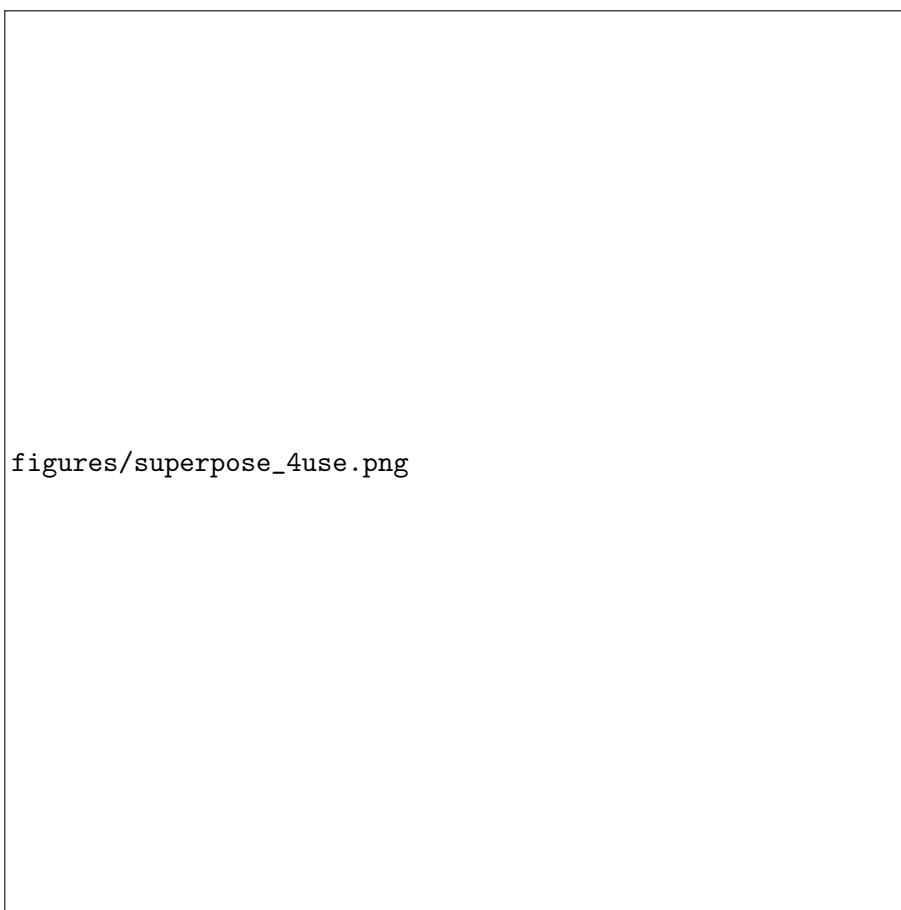
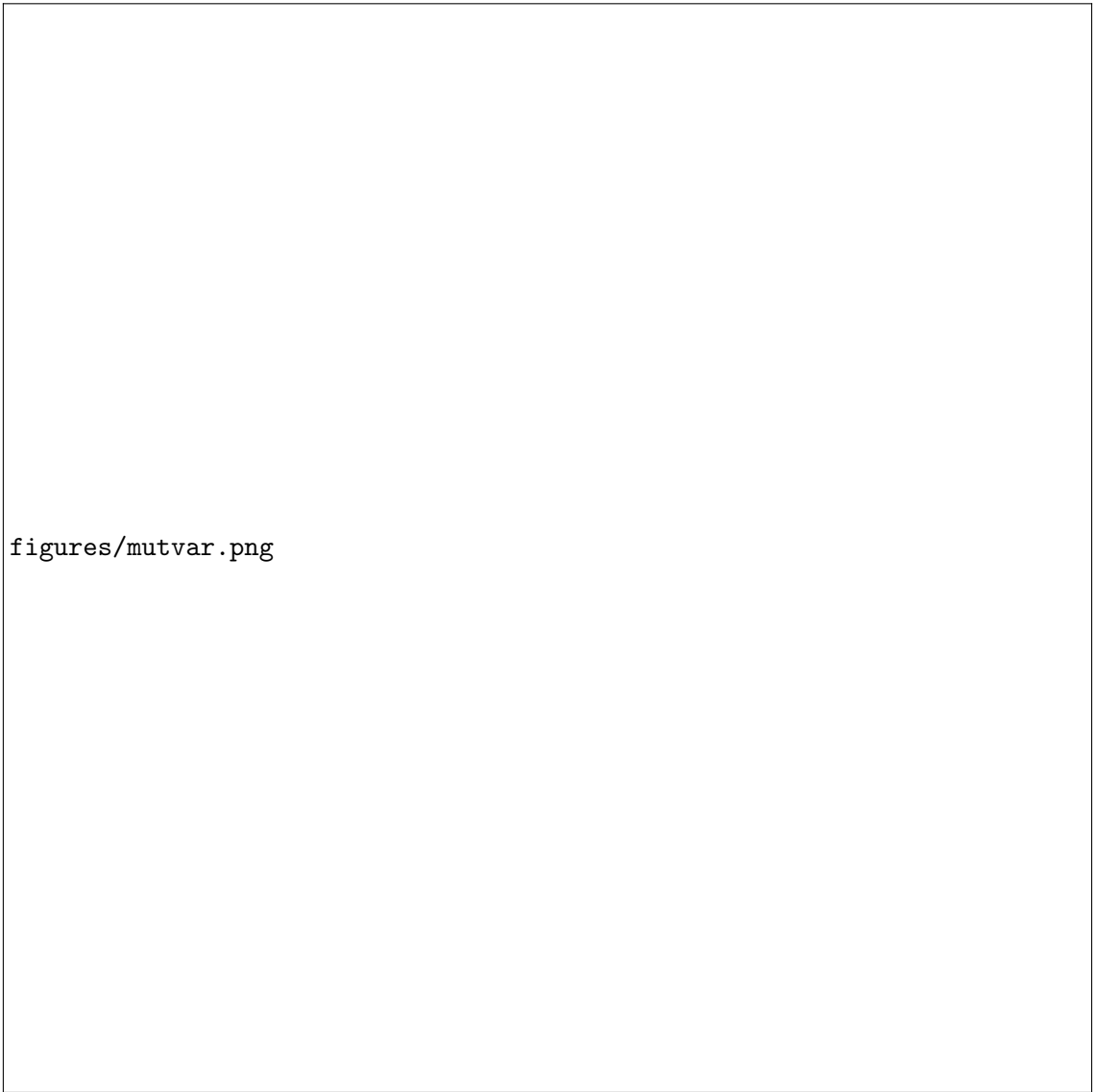


Figure 24: Structural alignment and superposition of the relatives in Pfam-FunFam based on the alignment of the binding region. The interacting residues are coloured in yellow while the secondary structures are coloured accordingly (Beta sheet(blue), alpha (Magenta), the inhibitor is coloured in rainbow.

### **Network analysis of MutFams enriched with kinases**

We considered the network properties of mutation families (MutFams) which are enriched domain families that are highly mutated in a disease condition (cancer in this case) and compare the outcome with human variation data (HumVar). This approach was taken to study the kinases involved in diseases and examine the network properties of these mutation families in comparison with the targeted kinase FunFams.




figures/mutvar.png

Figure 25: DS-measure of the MutFam in comparison with the HUMVAR

Figure 25 shows a plot of the distribution of the DS-score between MutFams and HumVar. There is a significant difference between these two sets of genes as the MutFams are highly clustered in the human functional network as compared to the HumVar ( $P = 0.00645$ ). This indicate that the relatives of the MutFams are closely associated in the protein network and the MutFams therefore provides a reasonable annotation of disease genes with lower side effects anticipated when these families are targeted in a diseased condition.





figures/muttarg.png

Figure 26: DS-measure of the MutFam in comparison with the targeted kinases

The DS-score of the MutFams were compared with the targeted kinases (figure 26), these different sets of genes tend to be clustered in the same fashion as there is no difference in the observed DS-score of the MutFams and the targeted kinases ( $P = 0.0191$ ). This indicate that the targeted kinases shares similar network characteristics as the mutated sets of genes that are implicated in human diseases. These MutFams are potential therapeutic targets that could be harnessed and considered for therapeutic purposes.

Table 5: The mutfam classes and their representation in Pfam-FunFams with the similarity measure in the protein functional network

Cancer Types	CATH-FunFam	Pfam-FunFam	%overlap	No of Drugs	Kernel-sim	DS-score
LGG	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
LGG	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
BLCA	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
BRCA	3.30.200.20.FF2866					
BRCA	2.30.29.30.FF22238	PF00069.FF25168	14.2	4	0.319	1
BRCA	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
BRCA	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
COAD	3.30.200.20.FF64824					
COAD	1.20.120.330.FF23932					
COAD	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
COAD	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
COAD	1.10.510.10.FF79298	PF00069.FF62569	100		0.584709	1
COAD	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
COAD	1.10.510.10.FF78966	PF07714.FF3369	35.7	1	0.097	
COAD	1.10.510.10.FF79140	PF07714.FF13122	100		0.220333	1
GBM	1.10.510.10.FF79478	PF00069.FF61939	45.5		0.694	1
GBM	3.30.505.10.FF4305				0.509667	
GBM	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
GLI	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
GLI	1.10.510.10.FF79478	PF00069.FF61939	45.5		0.694	1
GLI	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.509667	1
GLI	3.30.505.10.FF4305					
GLI	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
KIRC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
LAML	1.10.510.10.FF78745	PF07714.FF13026	53.1	17	0.184	1.2
LIHC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
LIHC	3.30.60.20.FF5564	PF00069.FF62318	23.9	6	0.3411	1.17
LUAD	3.30.200.20.FF1240					
LUAD	3.30.200.20.FF64824					
LUAD	1.10.510.10.FF79008	PF00069.FF27817	46.2		0.541267	1
LUAD	1.10.510.10.FF79228	PF00069.FF62599	71.4		0.588333	1.5
LUSC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5
PAAD	1.10.510.10.FF78763	PF00069.FF62351	43.5	1	0.155692	1.5
READ	1.10.510.10.FF78531	PF07714.FF13154	47.1		0.846	1.2
READ	1.10.1070.11.FF1687	PF00454.FF1812	37.3		0.662619	1.5
READ	1.10.510.10.FF78946	PF00069.FF62345	43.8	5	0.4562	1.72
SKCM	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
THCA	1.10.510.10.FF78531	PF07714.FF13154	47.1	2	0.846	1.2
UCEC	1.25.40.70.FF2223	PF00454.FF1812	43.2		0.662619	1.5

## References