Defining the Active Conformation of Typical Protein Kinase Domains from Substrate-Bound PDB Structures Enables Active-State AlphaFold2 Models for All 437 Human Catalytic Protein Kinases

Joan Gizzio[1]

Bulat Faezov[1,2,3]

Qifang Xu[1]

Roland L. Dunbrack, Jr.[1*]

[1] Institute for Cancer Research

Fox Chase Cancer Center

Philadelphia PA 19111

USA

[2] Kazan Federal University, Kazan, Russian Federation

[3] Institute for Protein Design, University of Washington Seattle

***Correspondence: Roland.Dunbrack@fccc.edu**

Abstract

Humans have 437 catalytically competent protein kinase domains with the typical kinase fold, similar to the structure of Protein Kinase A (PKA). The active form of a kinase must satisfy requirements for binding ATP, magnesium, and substrate. From structural bioinformatics analysis of 248 crystal structures of 54 unique substrate-bound kinases, we derived structural criteria for the active form of typical protein kinases. We include well-known requirements on the DFG motif of the activation loop and the N-terminal domain salt bridge, but also on the positions of the N-terminal and C-terminal segments of the activation loop that must be placed appropriately to bind substrate. With these criteria, only 130 of the 437 human catalytic protein kinases (30%) are in the Protein Data Bank in their active form. Because the active forms of catalytic kinases are needed for understanding substrate specificity and the effects of mutations on catalytic activity in cancer and other diseases, we used AlphaFold2 to produce models of all 437 human protein kinases in the active form. This was accomplished with templates from the PDB that resemble substrate-bound structures, shallow multiple sequence alignments of orthologs and close paralogs of the query protein, and application of the active-kinase criteria to the output models. We selected models for each kinase based on intramolecular ipSAE scores of the activation loop residues of these models, demonstrating that the highest scoring models have the lowest or close to the lowest RMSD to 29 non-redundant substrate-bound structures in the PDB. A larger benchmark of 117 active kinase structures with solved activation loops in the PDB shows that 71% of the highest scoring AlphaFold2 models had backbone RMSD < 1.0 Å to the benchmark structures and 92% were within 2.0 Å. Models for all 437 catalytic kinases are available at https://dunbrack.fccc.edu/kincore/activemodels. We believe they may be useful for interpreting mutations leading to constitutive catalytic activity in cancer as well as for templates for modeling substrate and inhibitor binding for molecules which bind to the active state.

INTRODUCTION

Protein kinases regulate most cellular processes in eukaryotes. In humans, their dysregulation is often involved in disease and they are therefore often targets in drug development, especially in cancer [1]. A large majority of human protein kinases take on a common fold first determined by Susan Taylor and colleagues in 1991 [2], consisting of an N-terminal domain of five beta strands and the C-helix and a largely helical C-terminal domain. The residues involved in catalytic activity are contained in the catalytic and activation loops that form a pocket for ATP binding and a binding site for substrates. Humans have 480 genes which contain at least one full-length typical protein kinase domain; 13 of these have two kinase domains, for a total of 493 typical kinase domains [3]. Of these 493 kinase domains, 437 domains from 429 genes are likely catalytic kinases, participating in phosphorylation of Ser, Thr, or Tyr residues on proteins, and 56 are likely pseudokinases. As of December 1, 2025, in the PDB there are structures for 305 human typical kinase domains, of which 277 are catalytic kinases and 28 are pseudokinases [4].

Active and inactive conformations of typical kinases have been classified in several ways [5-12]. The active form is generally very similar across kinases because of the requirements of binding ATP, magnesium ions, and substrate, which impart constraints on the conformation of the activation loop [13]. Early in the history of structure determination of kinases, a classification of structures into “DFGin” and “DFGout” was described [14]. In DFGin structures, the Asp side chain of the DFG motif is "in" the ATP binding site, and the Phe side chain of the DFG motif is in a pocket adjacent to the C-helix of the N-terminal domain. In DFGout structures, the Asp side chain is "out" of the active site, and the Phe side chain is removed from the C-helix pocket, allowing for the binding of Type 2 inhibitors such as imatinib that span both the ATP site and the C-helix pocket [15].

We previously used the presence of bound ATP, magnesium ion, and a phosphorylated activation loop to identify a set of 24 “catalytically primed” structures of 12 different kinases in the PDB [10]. We found that in addition to being "DFGin," these structures possess specific backbone and side-chain dihedral angles for the DFG motif ("*BLAminus*"), comprising the backbone dihedral angles of the residue immediately preceding DFG as well as the D and F residues, and the side-chain χ_1 dihedral angle of the DFG-Phe residue. They also possess a previously characterized salt bridge between a conserved glutamic acid residue in the C-helix and a conserved lysine residue in beta strand 3 of the N-terminal domain [16]. These structures are often referred to as “C-helix-in.” Using these basic criteria for active kinases, only 195 of 437 catalytic typical human kinase domains (45%) are currently represented in the PDB with active structures.

Previous efforts [5-12] have not considered whether an observed experimental structure can bind substrate, which may require additional structural features [13, 17]. In this paper, we present a structural bioinformatics analysis of 54 non-redundant substrate-kinase complexes from the PDB to define criteria for identifying catalytically active protein kinases structures, including the ability to bind substrate. In addition to the dihedral angle criteria on the X-DFG motif and a distance criterion on the N-terminal

domain salt bridge, we impose distance and dihedral angle criteria on the N and C terminal halves of the activation loop, whose positions are necessary for the formation of a substrate binding site. With these criteria, the number of human kinases with active structures is limited to 130 out of 437 catalytic protein kinases or 30%. Only 116 human catalytic protein kinases (26%) possess active structures and complete coordinates for the activation loop in the PDB.

The program AlphaFold2 from DeepMind is a deep-learning program for highly accurate protein structure prediction that is trained on structures from the PDB [18]. It takes as input the query sequence, a multiple sequence alignment (MSA) of homologs of the query, and optionally template structures related to the query. DeepMind has provided AlphaFold2 models of nearly all human proteins found in UniProt, which are available on a website provided by the European Bioinformatics Institute [19]. These were generated from a single structure produced by one seed of model 1.1.1 described in Jumper et al. [18]. However, with the criteria established in this paper, only 240 of the 437 (55%) catalytic human protein kinases have a fully active model in the EBI data set.

Because of the importance of knowing the active-state structures of kinases for understanding such features as substrate recognition, the effect of activating mutations in cancer, and drug development, we present a pipeline for producing active models of typical protein kinases using the program AlphaFold2. Several groups have found that using MSAs of reduced depth and templates in specific conformational states coerces AF2 to produce conformationally variable models, including some models in the conformational state of the templates [20, 21]. We use similar techniques to compute predicted structures of active kinases, using templates identified as active by our new criteria for active kinase structures, MSAs derived from orthologs of each kinase (with less than 90% sequence identity to each other), and applying the same criteria to the output models. From the active output models, we choose the top model using a modified intramolecular ipSAE score [22] based on aligned residues in the kinase domain and scored residues in the activation loops.

We benchmark our protocol with 29 substrate-bound kinase structures in the PDB with complete activation loops and a set of 117 kinase structures from the PDB that satisfy our active criteria and have complete (or nearly complete) activation loop coordinates around the substrate binding site. We show that the ipSAE scores for the activation loop are inversely correlated with RMSD of the activation loop for well characterized kinases. With these methods, we have produced active models of all 437 catalytic human protein kinase domains and made these models available on our Kinase Conformational Resource website and database, KinCore (https://dunbrack.fccc.edu/kincore/activemodels).

RESULTS

Catalytic protein kinases

We previously published an alignment of all 497 human kinase domains from 484 genes annotated by Uniprot [3]. This list excluded atypical kinases, such as ADCK, PI3/PI4, Alpha, FAST, and

RIO family kinases (https://www.uniprot.org/docs/pkinfam.txt). Since that time, three kinase genes have been identified as likely pseudogenes (SIK1B, PDPK2P, and PRKY) [23]. Omitting these pseudogenes and the truncated protein kinase domain in PLK5 (65 amino acids), we are left with 480 genes containing 493 typical protein kinase domains in the human proteome.

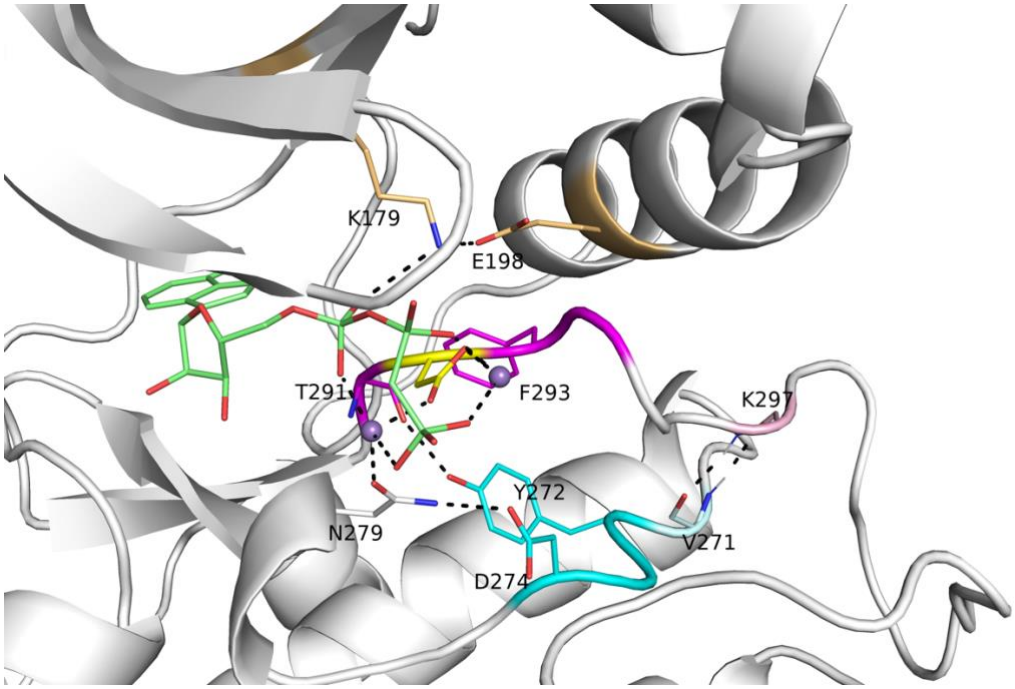


Figure 1. Active site of human AKT1 (PDB:4ekk, chain A).

Residues making hydrogen bonding interactions with ATP (green sticks), Mg^{2+} (purple spheres), the catalytic aspartic acid residue of the HRD motif (in AKT1, this is YRD; residues 272-274, cyan), and the aspartic residue (yellow) of the XDFG motif (residues 291-294, magenta) are shown in dashed lines. These include the salt-bridge residues of the N-terminal domain (K179, E198, gold). Residue K297, which is the sixth residue of the activation loop (light pink), makes backbone-backbone hydrogen bonds with V271, which immediately precedes the YRD motif. In the stick representations, oxygen atoms are in red and nitrogen atoms are in blue. A substrate peptide is present in this structure but not shown in this figure.

To make catalytically active models of all human kinases with typical protein kinase domains, we need to distinguish between catalytic protein kinase domains and non-catalytic protein kinase domains or pseudokinases. *Catalytic* protein kinase domains are those able to phosphorylate proteins on Ser, Thr, or Tyr residues. *Non-catalytic* protein kinase domains or pseudokinases are domains that possess the typical protein kinase fold but lack protein kinase activity, although they may have other catalytic activity (e.g., POMK). We identified catalytic protein kinases based on the presence of the Asp residue in the HRD motif, the Asp residue in the DFG motif, and the Lys residue of the N-terminal domain salt bridge. We reclassified WNK ("With No Lysine") kinases as catalytic protein kinases. Several kinases were reclassified based on literature annotations (e.g., BUB1B [24] and RYK are pseudokinases [25]). The result of these efforts was a list of 437 active kinase domains in 429 genes. Eight of these 429 genes have two (likely) catalytic protein kinase domains: RPS6KA1, RPS6KA2, RPS6KA3, RPS6KA4, RPS6KA5, RPS6KA6, OBSCN, and SPEG. Our previous phylogenetic analysis classified these 437 active domains into families as follows: AGC (60 kinases), CAMK (83), CK1 (11), CMGC (65), NEK (11),

OTHER (43), STE (45), TKL (37), and TYR (82). On our Kincore website (https://dunbrack.fccc.edu/kincore) and in this paper, we use the family name as a prefix in front of the HUGO gene name (e.g., TYR_BTK) [26]. The catalytic protein kinase domains and associated data are listed in **Supplementary Table 1**. The pseudokinase domains are listed in **Supplementary Table 2**.

Table 1. Kinase-substrate complexes in the Protein Data Bank (PDB)

Kinase	PDB	Substrate (Unip.)	Substrate Length	Auto	#Redundant Chains	Ligands	Dihedral label	Salt bridge
AGC_AKT1_HUMAN	4ekkA	GSK3B_HUMAN	10		12	ANP,MN	BLAminus	2.98
AGC_AKT2_HUMAN	1o6kA	GSK3B_HUMAN	10		13	ANP,MN	BLAminus	2.80
AGC_PRKACA_MOUSE	3x2uA	IPKA_HUMAN	20		11	ATP,MG	BLAminus	2.79
AGC_PRKACA_MOUSE	6mm5E	RYR2_MOUSE	12		3	ANP,MG	BLAminus	2.80
AGC_PRKACA_MOUSE	7e12A	PPLA_HUMAN	12		1	ANP,MG	BLAminus	3.27
AGC_PRKACA_MOUSE	3tnpC	KAP3_MOUSE	416		8	**None**	*ABAminus*	2.89
AGC_PRKACA_MOUSE	6mm7A	RYR2_MOUSE	209		3	ANP	ABAminus	**4.62**
AGC_PRKCI_HUMAN	5li1A	Q28E03_XENTR	20		2	ANP,MG	BLAminus	2.82
AGC_PRKCI_HUMAN	5lihA	KPCE_HUMAN	16		2	ADP,MN	BLAminus	3.03
CAMK_CAMK2A_HUMAN	7ujpA	KCC2A_HUMAN	22		6	ADP,MG	BLAminus	2.90
CAMK_CAMK2D_HUMAN	2welA	KCC2D_HUMAN	327	A	1	**K88**	BLAminus	**4.74**
CAMK_CAMKII_CAEEL	3kk8A	KCC2D_CAEEL	284	A	2	**MG**	BLAminus	**4.52**
CAMK_CAMKII_DROME	5fg8A	KCNAE_DROME	54		3	ADP,MG	BLAminus	2.80
CAMK_PHKG1_RABIT	2phkA	Peptide	7		1	ATP,MN	BLAminus	2.87
CAMK_PIM1_HUMAN	2bzkB	Peptide	15		35	ANP	BLAminus	2.73
CK1_CSNK1D_HUMAN	6ru7A	P63_HUMAN	16		6	ADP	BLAminus	2.80
CK1_CSNK1D_HUMAN	8d7nA	PER2_HUMAN	12		6	**None**	BLAminus	2.88
CMGC_CDK2_HUMAN	1qmzA	Peptide	7		5	ATP,MG	BLAminus	2.66
CMGC_CDK2_HUMAN	3qhrA	Peptide	10		4	ADP,MG	BLAminus	2.88
CMGC_DYRK1A_HUMAN	2wo6B	Peptide	8		1	**D15**	BLAminus	2.72
OTHER_CDC7_HUMAN	6ya7A	MCM2_HUMAN	15		1	ADP,ZN	BLAminus	2.77
OTHER_HASPIN_HUMAN	4oucA	H32_HUMAN	12		1	**5ID,IOD**	BLAminus	2.63
OTHER_PINK1_PEDHC	8uyhB	PINK1_PEDHC	463	A	24	ANP,MG	BLAminus	2.39
STE_PAK1_HUMAN	4zy4A	PAK1_HUMAN	301	A	6	**4T3**	BLAminus	3.90
STE_PAK4_HUMAN	2q0nA	Peptide	11		1	**None**	BLAminus	2.78
STE_PAK4_HUMAN	4jdhA	Peptide	15		2	**None**	BLAminus	3.32
STE_PAK4_HUMAN	6wlxA	CTNB1_HUMAN	7		1	**None**	BLAminus	3.14
STE_PAK4_HUMAN	6wlyA	LIMK1_HUMAN	10		1	**None**	BLAminus	3.05
STE_TNIK_HUMAN	7xzrB	Peptide	17		1	ANP,MG	BLAminus	2.58
TKL_BAK1_ARATH	3tl8A	HPAB2_PSESM	117		4	**None**	*ABAminus*	2.81
TKL_IRAK4_HUMAN	8wtfA	IRAK4_HUMAN	297	A	1	**8CG**	BLAminus	**4.63**
TKL_IRAK4_HUMAN	4u97A	IRAK4_HUMAN	312	A	1	**STU**	BLAminus	2.80
TYR_ABL1_HUMAN	2g2iA	Peptide	13		1	ADP	BLAminus	3.30
TYR_CSF1R_HUMAN	3lcdA	CSF1R_HUMAN	329	A	1	**BDY**	BLAminus	2.96
TYR_EGFR_HUMAN	4r3pA	ERRFI_HUMAN	7		2	**None**	BLAminus	**4.38**
TYR_EGFR_HUMAN	5czhA	Peptide	9		1	**None**	BLAminus	**3.78**
TYR_EGFR_HUMAN	5cziA	SHC1_HUMAN	9		1	**None**	BLAminus	**5.59**
TYR_EPHA2_HUMAN	4pdoA	EPHA2_HUMAN	299	A	2	**None**	None	**6.26**
TYR_EPHA3_HUMAN	3fxxA	Peptide	10		2	ANP	BLAminus	**4.25**
TYR_FES_HUMAN	3cd3A	Peptide	5		2	**STU**	BLAminus	2.45
TYR_FGFR1_HUMAN	3gqiA	FGFR1_HUMAN	326	A	1	ACP,MG	BLAminus	2.69
TYR_FGFR2_HUMAN	2pvfA	FGFR2_HUMAN	15	A	1	ACP,MG	BLAminus	2.77
TYR_FGFR2_HUMAN	3clyA	FGFR2_HUMAN	334	A	1	ACP,MG	BLAminus	2.62
TYR_FGFR2_HUMAN	5eg3A	PLCG1_RAT	114		1	ACP,MG	BLAminus	2.94
TYR_FGFR3_HUMAN	4k33A	FGFR3_HUMAN	325	A	3	ACP,MG	BLAminus	2.90
TYR_FGFR3_HUMAN	8uduB	FGFR3_HUMAN	297	A	1	**WIQ**	BLAminus	2.95
TYR_IGF1R_HUMAN	1k3aA	IRS1_HUMAN	14		1	ACP	*ABAminus*	**4.76**
TYR_IGF1R_HUMAN	3lvpC	IGF1R_HUMAN	336	A	1	**PDR**	*ABAminus*	**5.74**
TYR_INSR_HUMAN	1gagA	Peptide	13		3	**MG**	BLAminus	3.05
TYR_INSR_HUMAN	3bu5A	IRS2_MOUSE	15		5	ATP,MG	BLAminus	3.13
TYR_JAK2-2_HUMAN	6vnbA	JAK2_HUMAN	308	A	23	**R6P**	*ABAminus*	3.04
TYR_KIT_HUMAN	1pkgA	KIT_HUMAN	329	A	19	ADP,MG	BLAminus	**3.88**
TYR_PDGFRA_HUMAN	8pqhA	PGFRA_HUMAN	352	A	2	**9JI**	BLAminus	3.24
TYR_SYK_HUMAN	5c27A	Peptide	5		2	**50J**	BLAminus	2.79

Column 4 (substrate length) corresponds to the number of resolved Cα carbons in the substrate chain. Autophosphorylation complexes are marked with "A" in column 5. Outliers are shown in red (non-ATP structures, non-*BLAminus* structures, longer distances for some parameters). The absence of ATP is correlated with a broken salt bridge. Structures modeled as *ABAminus* in the PDB which are also supported by electron density (EDIA score of the XDFG carbonyl-O > 0.8) are colored red. Structures modeled as *ABAminus* with poor electron density support (XDFG carbonyl-O EDIA score < 0.8) are colored black and formatted in italic type (see [10]).

The characteristics of active protein kinase domains derived from substrate-bound structures

To identify structural features of the active form of catalytic protein kinases, we compiled a set of structures that constitute likely catalytically active structures of protein kinases. It consists of 248 structures of kinases in the August 2024 Protein Data Bank (PDB) with peptide or protein substrates bound at the active site (Supplementary Table S3); the set contains 54 unique kinase/substrate combinations (**Table 1**). The example shown in **Figure 1** is an active form of human AKT1 bound to a peptide substrate from GSK3B, PDB:4ekk [27]. Some of the "substrates" are in fact substrate-mimicking inhibitors, which bind similarly to substrates. Kinases in this table are represented more than once if: 1) they contain different bound substrates in the active site; 2) if the same substrate protein has multiple different constructs (e.g., peptide fragment vs full-length); or 3) if the same substrate protein is bound at different phosphorylation sites. There are twenty unique complexes with folded proteins domains, and seventeen of these are autophosphorylation complexes (marked "A" in column 5 of **Table 1**). Autophosphorylation complexes are homodimers of protein kinases inside crystals (or within cryo-EM structures) in which a known trans-autophosphorylation site of one monomer (Ser, Thr, or Tyr) is sitting in the active site of another monomer in the manner of a kinase substrate (i.e., similar to known peptide substrate complexes). We identified these as we did previously [28] by building out a 3x3x3 group of unit cells and searching for hydroxyl OH atoms of Ser, Thr, and Tyr of one monomer in hydrogen bonding distance of the aspartic acid side chain of the HRD motif of another monomer.

We previously identified three criteria for active structures in the PDB for catalytic protein kinase domains[10]: the spatial label must be *DFGin*; the dihedral label must be *BLAminus*; this indicates that the X, D, and F residues of the XDFG motif are in the "B", "L", and "A" regions of the Ramachandran map respectively, and the χ_1 rotamer of the Phe side chain is g^- (~ -60°); there must be a salt bridge between the C-helix glutamic acid side chain and the beta strand 3 lysine side chain (the WNK kinases are an exception to this rule). In **Table 1**, 47 of the 54 kinase-substrate pairs are in the DFGin-BLAminus conformation. As we noted previously, *ABAminus* structures involve a peptide flip [29] from *BLAminus* structures, and as much as 50% of these may be missolved [10]. Of the six *ABAminus* structures, only one of them has good electron density for the X residue's carbonyl oxygen (diagnostic of a mismodeled structure when the electron density is low). Twelve of 54 structures have a broken salt bridge but only two of these also have an ATP analog bound. All the structures are "Chelix-in".

We examined the 54 substrate-bound structures listed in **Table 1** for additional characteristics of the activation loop structure that may be required for binding substrates by determining contacts of the substrate with residues in the activation loop. These residues are in the appropriate position for forming an unobstructed substrate binding groove. Besides the conformation of the DFG motif and presence of the N-terminal domain salt bridge listed in **Table 1**, two other features are evident in the substrate-bound structures. The first is that the first few residues of the activation loop, up to at least the sixth residue (colored teal in **Figure 2**), have similar conformations and positions across all the structures, regardless

of family. The second is that the C-terminal segment of the activation loop, up to at least 10 residues from the end of the activation loop, also shares a common conformation and position across family members. This segment is sometimes referred to as the "P+1 loop," since it binds the side chain of the substrate residue immediately after the phosphorylation site [17, 30]. In **Figure 2**, residues 8 and 9 from the end of the activation loop (APE8 and APE9) are shown in dark orange. In non-TYR kinases, the conformation of residues 8-11 from the end of the activation loop resembles the hull shape of an upside-down, round-bottom boat. In TYR kinases, APE8 and APE9 are also in a common position, although the structure fluctuates in APE10 and APE11 far more than in the non-TYR kinase members.

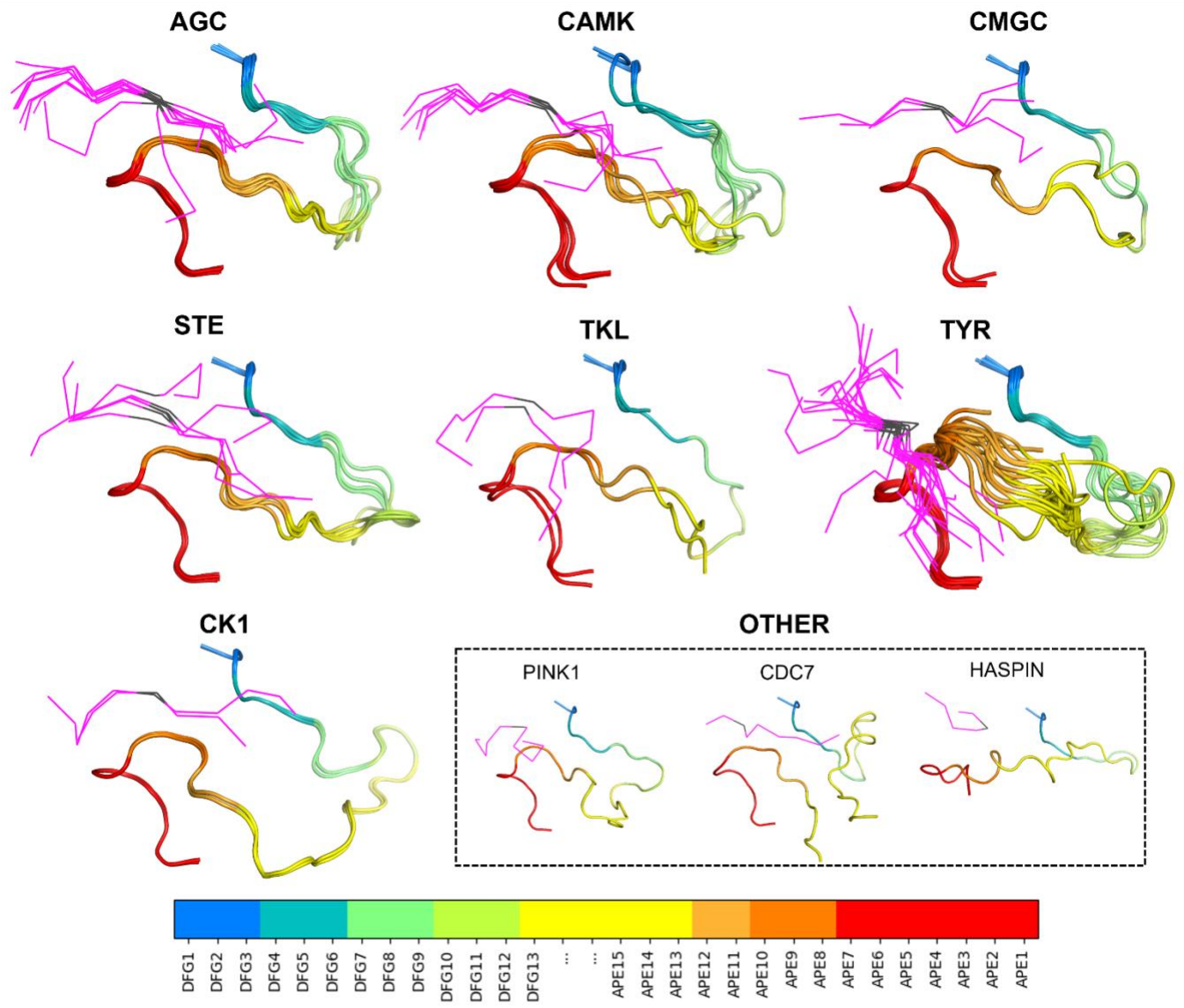


Figure 2. Substrate-bound structures from Table 1 organized by family.

In each figure, the substrate peptides (or pieces of longer proteins) are in magenta, the phosphoacceptor is colored black and the activation loop is colored according to residue number (see bottom legend). In the legend, N-terminal activation loop residues are labeled DFG*n* where *n* is the nth residue from the beginning of the loop, starting from DFG Asp (*n*=1) and ending at DFG13 (*n*=13). C-terminal activation loop residues are labeled APE*n* in reverse order where *n* is the nth residue from the end of the loop, starting from APE15 (*n*=15) and ending at APE1 Glu (*n*=1).

Below, we validate our previously defined active-state criteria: 1) DFGin; 2) BLAminus; 3) presence of salt bridge. Guided by the structures of substrate-bound kinases shown in **Figure 2**, we extend the criteria to include: 4) the activation loop must be "extended" as determined by the presence of a backbone-backbone hydrogen bond between the sixth residue of the activation loop (X in DFGxxX) and the residue before the HRD motif (X in XHRD); 5) the C-terminal segment of the activation loop must be positioned for binding a substrate, as determined by the conformation of residues six through twelve from the end of the activation loop (X in XXXXXXXxxAPE); 6) the backbone dihedral angles of the HRD motif must be in the α and L regions of the Ramachandran map for the His and Arg residues respectively. We also consider the presence of the regulatory spine defined by Kornev and Taylor [30]. We review each of these in turn.

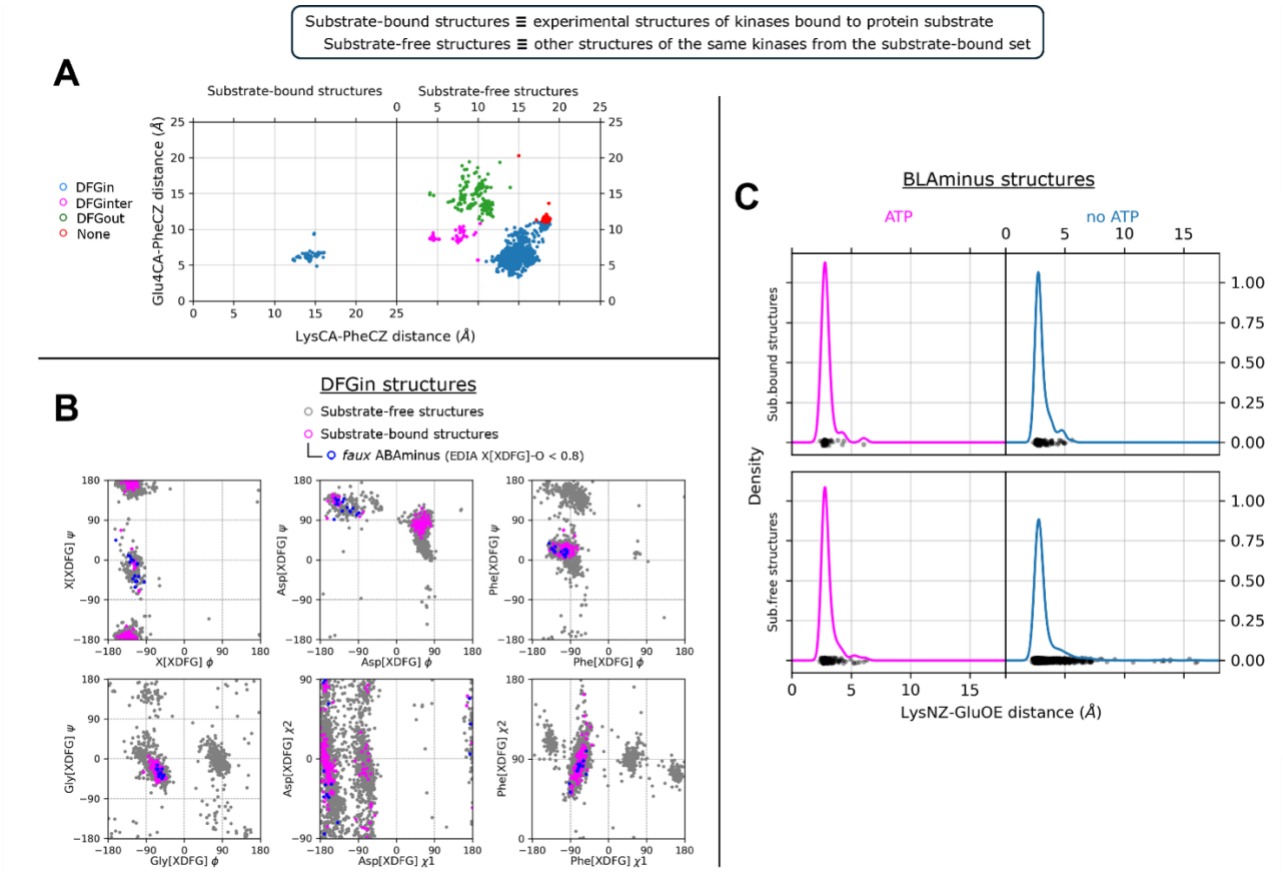


Figure 3. Standard measurements of kinase active state and the effects of ATP

A. Defining DFGin and DFGout structures. Distances of the Cζ atom of DFG-Phe from the Cα atoms of the salt bridge Lys residue and Glu4 (4 residues after the salt bridge Glu) for two different datasets: 248 substrate-bound structures (*left* panel), and substrate-free structures of the same kinases (*right* panel), as a control.

B. Ramachandran plots and side-chain dihedral angle plots for the XDFG motif residues of substrate-bound structures in the DFGin state (*magenta*) and substrate-free DFGin structures of the same kinases (*gray*). Substrate-bound structures modeled as *ABAminus* with poor electron density support (EDIA score of X[XDFG] carbonyl oxygen < 0.8) are colored blue.

C. Distribution of N-terminal domain Lys-Nζ (β3-strand) and Glu-Oε (αC-helix) distances for substrate-bound structures (*top* panels) and substrate-free structures of the same kinases (*bottom* panels) in the *BLAminus* state, comparing distances in structures bound to ATP analogs (*left* panels, magenta) vs those without ATP analog (*right* panels, blue). The presence of ATP analogs biases the formation of the Lys–Glu salt bridge. The presence of protein substrate biases the formation of this salt bridge even in the absence of ATP analogs. In the absence of substrate and an ATP analog, *BLAminus* alone cannot guarantee salt bridge formation.

DFGin conformation

The position of the DFG-Phe residue determines, in part, the position of the catalytic DFG-Asp residue. As we did previously [10], we define DFGin by the distance between the DFG Phe Cζ atom and the Cα atoms of two residues in the N-terminal domain: the Lys residue in the β3 strand of the N-terminal domain salt bridge and the "Glu4" residue in the C-helix (**Figure 1**), which is the residue four residues following the Glu residue of the salt bridge. Based on these distances, structures are labeled as follows: *DFGin*, where the DFG-Phe residue is near the C-helix Glu4 residue but far from the Lys residue; *DFGout*, where the Phe residue is far from the C-helix Glu4 residue and close to the Lys residue; and *DFGinter*, where the Phe residue is not far from either the Glu4 or Lys residues. These distances are plotted for substrate-bound and a *control group* consisting of substrate-free structures of the same kinases in **Figure 3A**. All 248 substrate-bound structures identified in the PDB are DFGin (required for the *BLAminus* and *ABAminus* conformations of the XDF motif).

BLAminus conformation and Salt bridge formation

The conformation of the XDFG motif and the formation of the salt bridge in the N-terminal domain work together to form an active site capable of binding ATP and magnesium ions for the phosphorylation reaction. These interactions are shown in **Figure 1**, where the Asp of the DFG motif interacts with the active site magnesium ions which chelate ATP. The carbonyl oxygen of the residue before the DFG motif (X of XDFG, T291) forms a hydrogen bond with the Tyr residue of the YRD motif in AKT1 (usually HRD, but YRD in AKT1). This hydrogen bond helps position the catalytic aspartic acid residue of the Y/HRD motif, which interacts with the Ser, Thr or Tyr hydroxyl atoms of substrate residues to be phosphorylated. The *BLAminus* conformation is required for these interactions [10]. As noted above, *ABAminus* structures involve a "peptide flip" of the X-D residues [29], such that the carbonyl of the X residue points upwards and does not interact with Y/H of the Y/HRD motif (due to a 180° change in ψ at the X residue of XDFG and a 180° change in ϕ of the Asp). Many of these structures are missolved and should be *BLAminus*, as demonstrated by poor electron density for the X residue carbonyl oxygen [10, 31]. Of the 248 substrate-bound structures, 34 of them are *ABAminus*: 18 of these 34 have poor electron density of the X residue's carbonyl oxygen, according to the EDIA (Electron Density score for Individual Atoms) program [32]. The backbone and side-chain dihedral angles of the XDFG motif for all substrate-bound structures identified in the PDB (including those listed in **Table 1**) are plotted in **Figure 3B**.

The Lys of the N-terminal domain salt bridge interacts directly with the alpha-beta phosphate linkage of ATP. The Glu of the salt bridge helps position the Lys in this interaction [16]. The *g*-minus rotamer of the Phe side chain is required for this interaction, since the *g*+ rotamer of the inactive *DFGin-BLAplus* and *DFGin-BLBplus* conformations points upwards (instead of downwards as in the *BLAminus* and *BLBminus* conformations) and pushes the C-helix outwards, breaking the salt bridge [10]. In **Figure 3C** the distribution of distances of the salt bridge atom pairs (Nζ in the β3 Lys residue with $O\varepsilon_1$ or $O\varepsilon_2$ in

the C-helix Glu residue, whichever is shorter) is compared for *BLAminus* structures with/without ATP and with/without protein substrate. When *BLAminus* structures are bound with ATP, the salt bridge is strongly favored with a mean distance of about 3.0 Å regardless of whether protein substrate is also bound (left panels of **Figure 3C**). However, in the absence of ATP and protein substrate there is a significant number of *BLAminus* structures with broken salt bridges. From this we emphasize that the *BLAminus* state alone is an insufficient proxy for catalytic potential, and the salt bridge must be measured as well to determine whether a structure would form catalytically competent interactions with ATP.

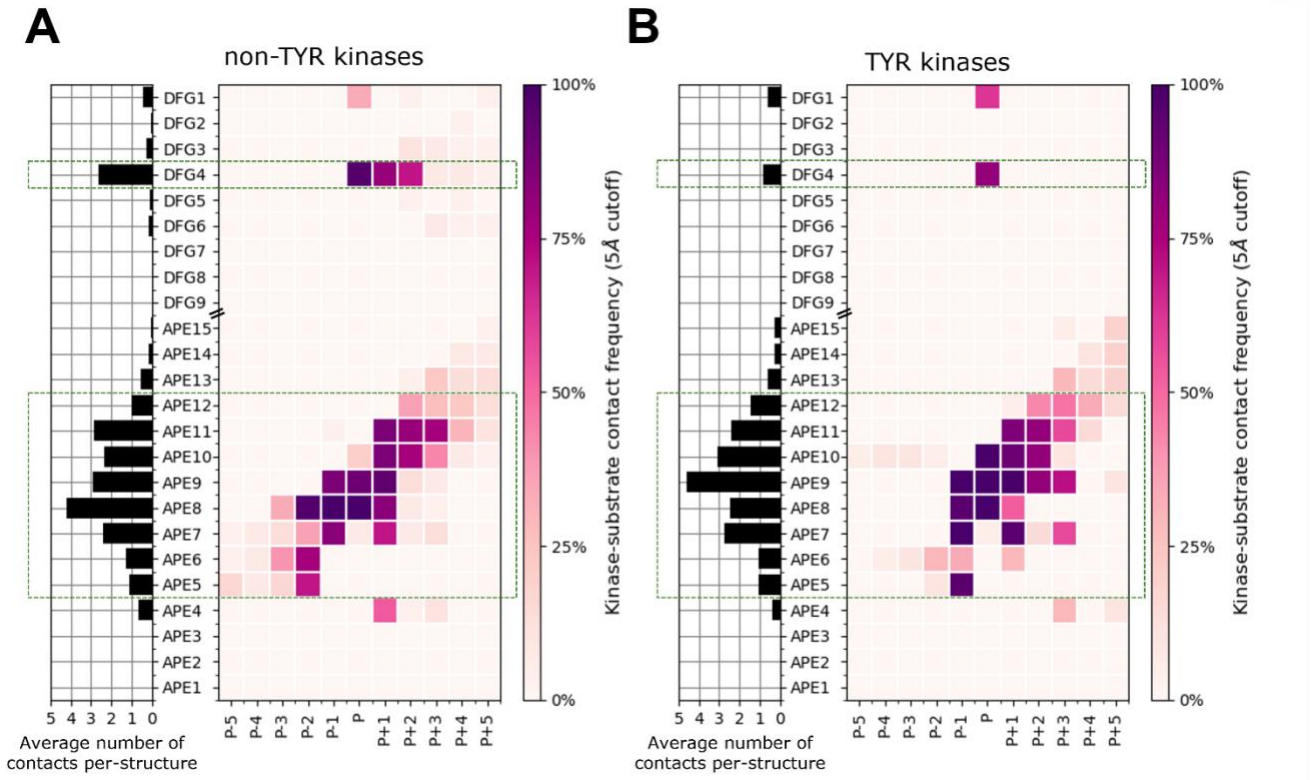


Figure 4. Substrate orientation and activation loop contacts for all 248 substrate-bound kinases in the PDB.

N-terminal activation loop residues are labeled DFG*n* where *n* is the nth residue from the beginning of the loop, starting from DFG Asp (*n*=1) and ending at DFG9 (*n*=9). C-terminal activation loop residues are labeled APE*n* in reverse order where *n* is the *n*th residue from the end of the loop, starting from APE15 (*n*=15) and ending at APE1 Glu (*n*=1). Substrate residues are labeled P-5 through P+5 where P is the substrate phosphoacceptor. Activation loop residues that form contacts with approximately one or more substrate residues per structure (on average) are highlighted with dashed boxes. Left-side panel in **A** and **B**: Average number of substrate contacts formed by each activation loop residue per-structure, which is ≥1 for APE12-APE5.

A. Substrate contacts with the activation loop in non-TYR-family kinases. Left panel: Average number of substrate contacts formed by each activation loop residue per-structure, which is ≥1 for DFG4 and APE12-APE5.

B. Same as (A) but for TYR-family kinases, which bind substrates in an overall different orientation than non-TYR-family kinases.

ActLoopNT and *ActLoopCT conformations required for substrate binding*

To investigate potential requirements for substrate binding, we determined which residues in the activation loop form direct contacts with substrate residues (any atom-atom contact within 5 Å between substrate residues and the DFG…APE sequence). The results for unique complexes from **Table 1** are shown in **Figure 4** and listed in **Supplementary Table 4**. In **Figure 4**, we plotted the frequency of contact

of substrate residues P-5 to P+5 (where P0 is the phosphorylation site itself) with activation loop residues counting from the beginning of the activation loop (DFG1=Asp, DFG2=Phe, DFG3=Gly, etc.) and from the end of the activation loop (APE1=Glu, APE2=Pro, APE3=Ala, etc.). The results are shown in **Figure 4A** and **Figure 4B** for non-TYR and TYR family kinases respectively.

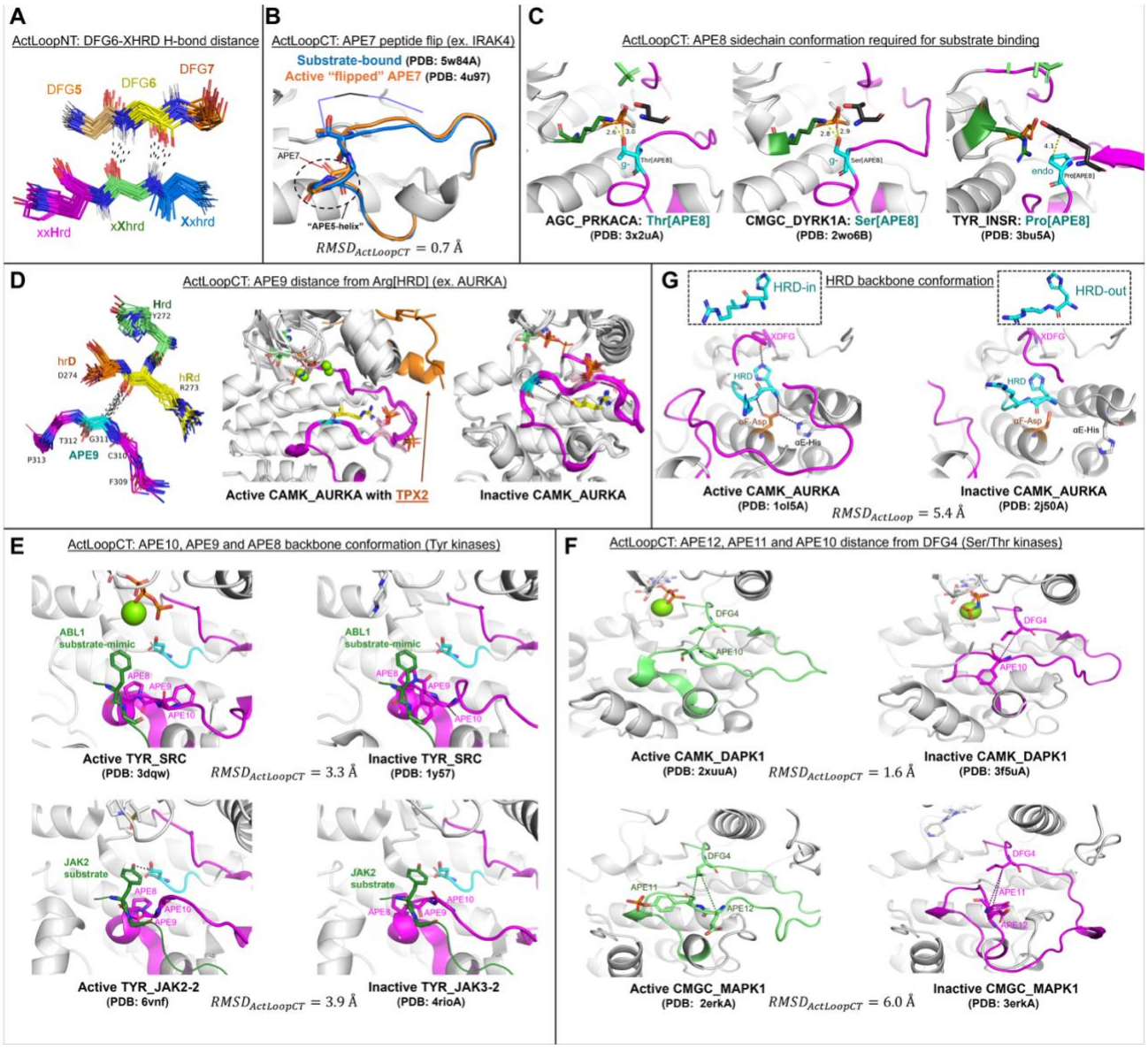


Figure 5. Conformational states involving the activation loop that correlate with the binding of protein substrates

A. Beta bridge hydrogen bonds between DFG6 and XHRD residues in kinase-substrate complex structures listed in Table 1. The carbon atoms are colored as follows: DFG5 (gold), DFG6 (yellow), DFG7 (orange), Xxhrd (blue), xXhrd (green), xxHrd (magenta, including the side chain, which is sometimes Tyr). Oxygen atoms are in red, hydrogen atoms are in white (modeled with PyMol), and nitrogen atoms are in blue. Hydrogen bonds in a few selected structures are marked with dashes.

B. Example of an APE7-APE6 peptide flip in IRAK4 where both the flipped and un-flipped structures are considered active by our criteria. The differences in global ActLoopCT conformation between the two structures is very small (RMSD ≈ 0.7 Å). The active substrate-bound structure of IRAK4 and its substrate are colored blue, and the active peptide-flipped structure is colored orange.

C. Non-TYR-family kinases typically have Thr at APE8 (*left*) or Ser (*center*) and the g^- rotamer of these sidechains allows APE8 to form simultaneous hydrogen bonds with the conserved Asp and Lys of the HRD loop (HR**D**x**K**). Tyr kinases have Pro at APE8 (*right*) where the gamma carbon forms a short contact with Tyr substrates when it is in the Cγ-*endo* rotamer state (χ_1>0°).

Transition to Cγ-*exo* (χ_1<0°) would likely inhibit catalysis as described in the main text.

D. (*left* panel) Contact between Cα of the APE9 residue (XxxxxxAPE) and backbone carbonyl oxygen of HRD2 (conserved Arg of H**R**D motif). His/Tyr (green), Arg (yellow), and Asp (orange) of the HRD loop are shown in sticks, including side chains, numbered according to AKT1 residues 272-274. Residues APE11 (AKT1 F309, magenta), APE10 (C310, magenta), APE9 (G311, cyan), APE8 (T312, magenta), APE7 (P313, magenta) are shown in sticks without side chains. These structures correspond to PDB entries from the AGC, CAMK, CMGC, STE, and TKL kinases listed in Table 1 of the main text. (*middle* and *right* panels) Two conformations of human AURKA. (*middle* panel) Active structures with bound TPX2 (orange): PDB: 1ol5, 3e5a, 3ha6, 5lxm, 6vph. (*right* panel) Inactive structures without TPX2. PDB: 4dee, 5dt3, 5oro, 5os0, 6i2u, 6r49, 6r4d and others. Gly291 Cα (cyan sticks) is in contact with HRD-Arg255 backbone carbonyl O (yellow sticks) in the active structures (average distance 3.6 Å), while there is no contact in the inactive structures (average distance > 10 Å). The activation loop C-terminal region in the active structures resembles the structures of substrate-bound complexes, while the inactive structures would block substrate binding.

E. Comparison of active (*left* panels) TYR-family kinases and examples from the same family which are inactive (*right* panels) due to substrate-incompatible ActLoopCT dihedrals, superimposed with a bound-substrate from the most closely related family member. In the upper panels the substrate-mimic (P0 Tyr → Phe mutant) from TYR_ABL1 (PDB: 2g2i) is superimposed with active and inactive (ActLoopCT-out) structures of TYR_SRC. In the bottom panels the substrate from active TYR_JAK2 (green) is shown for reference.

F. Comparisons of active (*left*: green) and inactive ActLoopCT-out structures (*right*: magenta) of CAMK_DAPK1 (*top row*) and CMGC_MAPK1 (*bottom rows*) (also called ERK2). The inactive DAPK1 structure (3f5uA) has an expanded APE10-DFG4 distance relative to the active conformation (2xuuA). The inactive MAPK1 structures have a contracted APE11-DFG4 distance (4xj0A) and expanded APE12-DFG4 distance (3erkA) relative to the active, phosphorylated conformation (2erkA).

G. Structural visualization of CAMK_AURKA in the active (*left*) vs inactive/HRD-out (*right*) state. The HRD-out conformation is a peptide flip relative to HRD-in that breaks a hydrogen bond between the HRD-Arg backbone nitrogen and the conserved Asp in the αF-helix (orange). These hydrogen bonds stabilize a salt bridge between the αF-Asp and conserved His in the αE-helix. The inactive DFGinter structure of CAMK_AURKA, 2j50A (right panel) demonstrates how the HRD peptide flip can propagate structural changes to the XDFG motif and αE-His, resulting in a more globally inactive conformation.

TYR-family kinases bind substrates in an overall different orientation from non-TYR kinases. Relative to non-TYR kinases, TYR kinases lack contacts between P+1/P+2 and DFG4 and between P-2 and APE8 and gain contacts between P-1 and APE5, between P and APE10, and between P+2 and P+3 with APE9. For both families, contacts with the substrate P-sites and flanking residues occur mainly with the N- and C-terminal segments of the activation loop. We focus our analysis on these two segments to arrive at a set of conformational criteria for active kinases beyond what is currently described in the literature. Structural depictions of these criteria and examples of "pseudo-active" structures that violate them (structures which we now designate as "inactive") are depicted in **Figure 5** and referenced in the following sections.

ActLoopNT. We find that most substrates have a contact with one or more of the DFG-adjacent residues and, notably, residue 4 of the activation loop (DFG4). For all substrate-bound structures, the DFG4 residue interacts with the phosphoacceptor residue ("P-site"). For non-TYR kinases, due to the different orientation of the substrate in these kinases, the DFG4 residue is also in contact with the P+1 residue, and in a smaller number of cases, forms contacts with P+2. Based on these observations, it is not surprising to find that, in the activation loops of active kinase structures, DFG4 (DFG**X**xx) is flanked by two structural motifs that constrain its positioning: besides DFG-in/*BLAminus* (**XDF**Gxxx), there are two anti-parallel beta strands formed by backbone-backbone hydrogen bonds between residue 6 of the activation loop (DFGxx***X***) and the residue immediately preceding the HRD motif (***X***hrd) (**Figure 5A**).

These short beta strands (3 residues each, centered on DFG6 and Xhrd) have been labeled beta strand 6 (comprising the XXH residues of XXHrd) and beta strand 9 (comprising residues 5-7 of the activation loop, dfgxXXX) in protein kinase structures [30]. This hydrogen bond is present in all of the substrate-bound structures in **Table 1** ("DFG6" in the table) and is likely an evolved feature of the active conformation that positions DFG4 to interact with substrate P-site and P+1 residues.

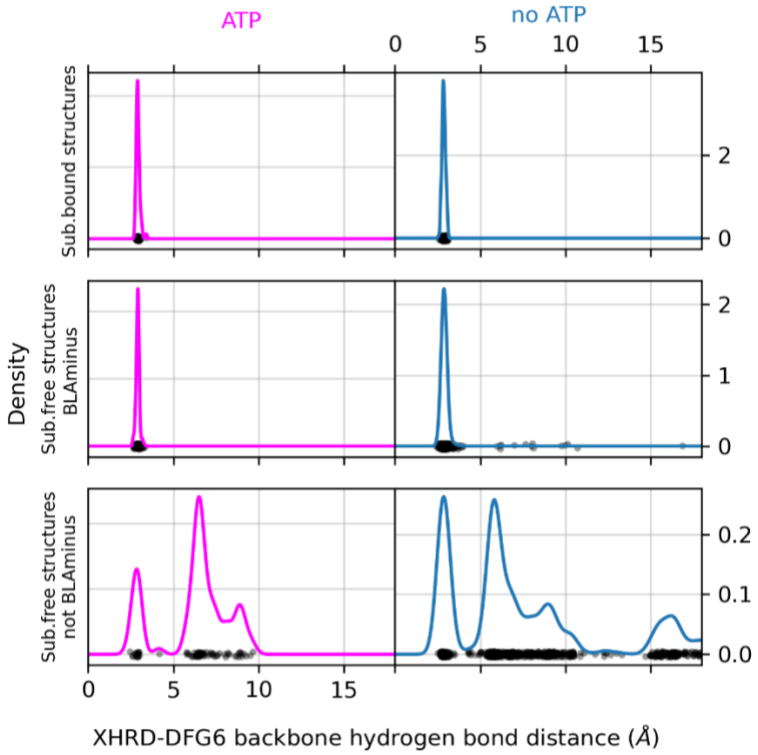


Figure 6. Interactions of residue 6 of the activation loop and the residue before the HRD motif ("XHRD")

Distribution of the XHRD-DFG6 backbone-backbone hydrogen bond distance in substrate-bound structures (*top* panels), substrate-free structures which are also *BLAminus* (*middle* panels), and substrate-free structures which are *not BLAminus* (*bottom* panels), comparing structures bound to ATP analog (*left* panels, magenta) vs those which are not bound to ATP analog (*right* panels, blue). The distance plotted is the minimum of the N-O or O-N distances between these two residues. The DFG6 residue is identified with the last X in the DFGxxX sequence, where x/X is any amino acid. The XHRD-DFG6 hydrogen bond is formed (ActLoopNT-in) in all substrate-bound structures regardless of whether an ATP analog is present (*top* panels). In the control structures without protein substrate and without ATP analog (*middle* and *bottom* panels) the distance distributions adopt a wider range. Structures that are not *BLAminus* are mostly ActLoopNT-out.

This DFG6/Xhrd hydrogen bonds also cooperate with the conformation of the XDF motif: 99% of *BLAminus* structures (660 of 666) with bound ATP or ATP analogs contain this hydrogen bond with the minimum N-O or O-N backbone-backbone distance less than 3.6 Å (**Figure 6**, upper and middle panels). The only exceptions are the partially disordered activation loop structures of the bacterial kinase PknB (4eqm chains D and E) and TYR_HCK (1ad5 chains A and B), and an activation-loop swapped structure of STE_MAP4K1 (6cqd chains A and B) in which the N-terminal portion of the activation loop forms an α-helix. Without ATP and protein substrate, 2% of *BLAminus* structures have a broken DFG6/XHRD hydrogen bond (**Figure 6**, middle right panel). Although this fraction is small (comprising ~150 PDB chains), it indicates that the DFG6 criterion is useful for distinguishing active conformations among

BLAminus structures. When the XDF state is not *BLAminus* the fraction of structures with a broken DFG6/XHRD hydrogen bond increases from 2% to 75% (**Figure 6**, bottom right panel).

ActLoopCT. The conformation of the C-terminal end of the activation loop is critical for binding substrates. Most substrate-bound structures in **Table 1** contain contacts between the substrate and residues 4-12 from the end of the activation loop (detailed in **Supplementary Table 4**), which usually ends in the sequence motif "APE" (APE in 306 kinase domains, SPE in 47, PPE in 29, and other sequences in the remaining 55 kinases). The positioning of the APE motif is conserved in most catalytic kinases, owing to its close packing with the αF-helix.

Differences in the bound orientation of substrates for the TYR family compared to non-TYR families require the C-terminal region of the activation loop to adopt a different, more distended conformation (**Figure 2**). This results in distinct sets of contacts with the substrate (**Figure 4**), suggesting that TYR family kinases require different conformational criteria of the activation loop C-terminal region to accurately define the active conformation. In TYR kinases, the substrate binds directly to these residues in the form of a short beta strand (magenta lines in **Figure 2**). In other kinase families, the substrate binds to a groove between the N and C-terminal segments of the activation loop.

*APE5 helix***.** Moving in the N-terminal direction from the APE motif, APE5-APE7 belong to a short helix-like structure that interacts with substrate residues P-1 through P-5 (**Figure 4**). The APE5 helix can be captured by imposing conformational criteria on the backbone dihedrals ϕ and ψ of APE7 and APE6, requiring both residues to fall in the broad alpha helical "A" region of the Ramachandran map: $\phi \in (-180°,0°)$, $\psi \in (-100°,50°)$. In addition to interacting directly with substrate residues P-3 through P-1 (**Figure 4**), the APE5 helix also constrains the conformational variability of the rest of the C-terminal segment (APE8 – APE12) allowing these residues to form an extended binding surface for the substrate as described below.

In a few structures, there is a peptide flip, such that residues APE7 and APE6 fall in the B/beta and L/left-handed regions respectively, preserving the trajectory of the backbone structure (**Figure 7A-B**). For example, residues APE7 and APE6 of TKL_IRAK4 (5w84A) appear in the B-L Ramachandran regions (**Figure 7A**, left panel), and structural superposition with the substrate-bound structure of IRAK4 (4u97A) reveals that the main difference between the two structures is the orientation of the APE7 carbonyl oxygen, which is flipped "up" in the former (B-L) and pointing "down" in the latter (A-A) (**Figure 5B**). The rest of the ActLoopCT conformation is largely unaffected by the flip (ActLoopCT RMSD < 1Å). The structural change involves a flip from a Type I beta turn (A-A) to a Type II beta turn (B-L) [33]. In the active-kinase criteria, we allow for this peptide flip (see below), such that for APE7: $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$ and for APE6: $\phi \in (0°,180°)$, $\psi \in (-50°,100°)$. Most of the kinases that contain the flip have a glycine residue at position 6 (AGC_GRK1-7, TKL_LRRK1-2, OTHER_ERN1-2), which is unsurprising given the L conformation.

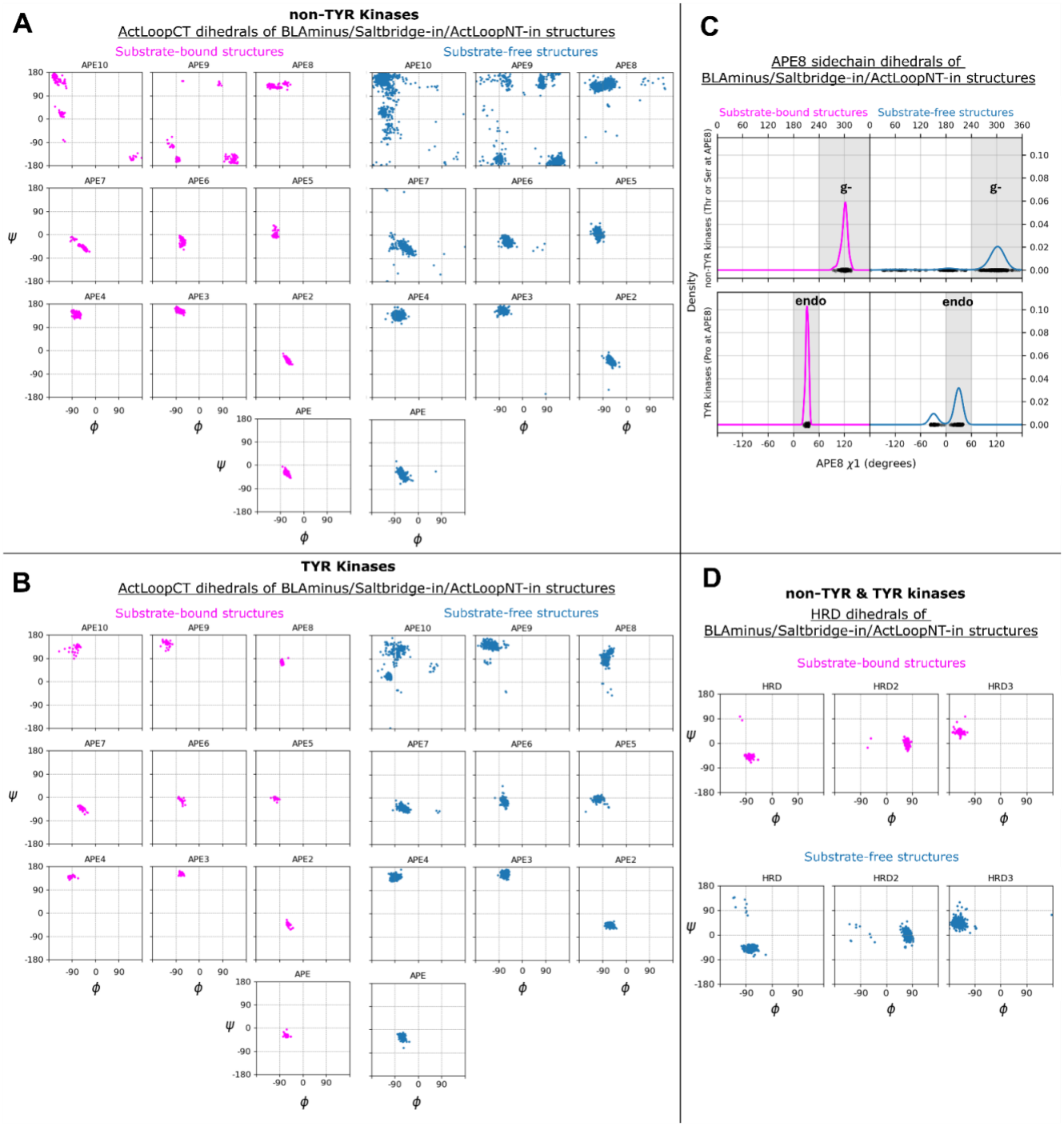


Figure 7. Dihedral criteria for Activation Loop C-terminus and catalytic HRD loop residues

Comparison of 248 substrate-bound structures vs substrate-free structures of the same kinases (as controls) which satisfy previously defined active-state criteria (*BLAminus*, Saltbr-in, ActLoopNT-in) to identify the need for active-state criteria based on dihedral angles.

A. ActLoopCT backbone dihedrals for substrate-bound non-TYR kinase structures (magenta, *left*) compared with the substrate-free set comprised of the same kinases (blue, *right*).

B. ActLoopCT backbone dihedrals for substrate-bound TYR kinase structures (magenta, *left*) compared with the substrate-free control set comprised of the same kinases (blue, *right*).

C. APE8 sidechain χ_1 dihedrals for non-TYR kinases (Ser[APE8] and Thr[APE8]) and TYR kinases (Pro[APE8]) compared between substrate-bound (*left,* magenta) and substrate-free (*right*, blue) sets. Ser[APE8] and Thr[APE8] side chains in the non-TYR kinase substrate-bound set are always in the g- conformation. Pro[APE8] sidechains in TYR kinase substrate-bound set

(*bottom left* panel) are always Cγ *endo* (χ_1>0°). APE8 side chains in the substrate-free sets can deviate significantly from these conformations.

D. HRD loop backbone dihedrals for the substrate-bound (*top*) and substrate-free (*bottom*) sets plotted for both non-TYR and TYR kinases. Peptide flips at the HRD-His and HRD-Arg positions (conserved His and Arg, respectfully) can be observed in the substrate-free set which has functional consequences for kinase activity.

APE8. For both TYR and non-TYR kinases, the APE8 residue is critical for binding substrates, interacting with substrate residues P-2 to P+1 with high frequency in non-TYR kinases (**Figure 4A**) and with P-1 to P+1 in TYR Kinases (**Figure 4B**) as well as catalytic residues in the HRD loop. APE8 is a conserved Pro in all 83 TYR kinases, and Ser or Thr in 335 non-TYR kinases (95%). Atypical residue types such as Asn can also be observed at this position (e.g., OTHER_WEE1 and OTHER_WEE2). We also note that the APE motif and its helical structure are missing altogether from three unusual human kinases (OTHER_HASPIN (4oucA from **Table 1**), OTHER_PKDCC, OTHER_TP53RK) which do not have an APE motif [3]. None of the APE criteria (above and below) are applied to these three kinases.

For non-TYR kinases, all substrate-bound structures have a backbone conformation in the beta region of the Ramachandran map (**Figure 7A**), while non-substrate bound structures in the *BLAminus* conformation are sometimes in the upper right of the Ramachandran map (top right of **Figure 7A**). Flipping ϕ of APE8 (assuming APE7 and APE6 are properly placed) would break typical backbone-backbone hydrogen bonds of APE9 with P+1 of the substrate. We therefore place a requirement on the ϕ,ψ values of APE8 of non-TYR kinases: $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$.

When APE8 is Ser or Thr in non-TYR kinases, it is always found in a gauche-minus rotamer (-60°). The g^- rotamer orients the gamma hydroxyl to accept a hydrogen bond from the side-chain Nζ of a conserved Lys in the HRD loop (HRDx**K**xxx); simultaneously it donates a hydrogen bond to the side-chain carboxylate of the conserved Asp in the HR**D** motif (**Figure 7C,** top panel and middle panel) (**Figure 5C**). These hydrogen bonds are present in all but three of the substrate-bound kinases from **Table 1** that contain Thr/Ser at APE8 (not counting 5lihA which is missing atoms from Lys[HRDx**K**]). This configuration promotes catalysis by orienting the other carboxylate oxygen atom "upward" toward the substrate hydroxyl. Consequently, we place constraints on the χ_1 rotamer of the APE8 residue: when this residue is a Ser or Thr in non-TYR kinases, we require the g- rotamer: $\chi_1 \in (-120°, 0°)$.

In TYR family kinases, APE8 is instead Pro. In substrate-bound structures, the backbone conformation is always close to (-75°,+85°) (**Figure 7A**), and the side-chain conformation is always in a Cγ-*endo* rotamer (χ_1 ~ +30°) (**Figure 7C**, bottom panel). This is consistent with data from the PDB, where nearly all Pro residues with ψ between 45° and 135° are in the Cγ-*endo* rotamer and their ϕ dihedral is < -70° [34]. This proline is in the middle of an inverse gamma turn with a hydrogen bond between the carbonyl oxygen of APE9 and the backbone NH of APE7 (NB: APE7 is a proline in 45% of non-TYR human kinases [3]). The Cγ-*endo* rotamer creates a small divot in the proline ring, creating space for the Tyr substrate which forms a short contact with the Pro beta carbon (**Figure 7C**) (**Figure 5C**); the Cγ-*exo* rotamer would lift this gamma carbon by ~1.4 Å, pushing the substrate away from its hydrogen bond

acceptor and donor (Asp and Arg of HR**D**xxx**R**). The rotamer of proline is strongly correlated with the backbone dihedrals ϕ and ψ with a small energy barrier of less than 3 kcal/mol between the Cγ-endo and Cγ-exo conformations [35, 36]. Consequently, we do not put a requirement on the side-chain conformation of this residue for TYR kinases, but we do put the same restriction on the backbone dihedrals as for non-TYR kinases: $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$. The Pro residue at position APE8 enforces a beta conformation on the APE9 residue (see below), which is therefore a pre-Pro residue, whose Ramachandran map is distinct from non-pre-Pro residues [37, 38], with significantly reduced populations in the alpha region.

APE9. From examination of the substrate-bound structures, we identified a contact that is consistent with substrate binding, and which is absent in structures that likely block substrate binding: a contact (or near contact) between the APE9 Cα atom and the backbone carbonyl oxygen of the Arg residue in the HRD motif (HRD-Arg). **Figure 5D** shows this contact for 23 non-TYR kinase structures from Table 1. The Cα-O distance is ≤ 4.2 Å in all of these structures.

Aurora A kinase (CAMK_AURKA) is a good example of the utility of the APE9 distance criterion. In the ATP-bound, *BLAminus* state, there are two dominant conformations of the entire activation loop of AURKA, one where the APE9/HRD-Arg contact is formed and the other where the contact is broken. **Figure 5D** shows five structures that contain these contacts. This comprises five structures of AURKA with TPX2 (PDB: 1ol5, 3e5a, 3ha6, 5lxm, 6vpg). Two other structures bound with MYCN (PDB: 5g1x, 7ztl) are very similar (not shown). Both proteins are known to activate AURKA by binding to the N-terminal domain and the tip of the activation loop [39, 40]. Most *BLAminus* structures of AURKA, however, resemble the structures shown in the right panel of **Figure 5D** where the APE9-Arg contact is broken (e.g., PDB: 4dee, 5dt3, 5oro, 5os0, 6i2u, 6r49, 6r4d). In these structures, the C-terminal end of the activation loop (APE6-APE10) deviates significantly from the TPX2- and MYCN-bound structures and from the structures of substrate bound kinases in the related AGC and CAMK families. The Cα-O distances in these AURKA structures are all more than 10 Å.

We examined the distributions of this distance (**Figure 8A**) for both substrate-bound and substrate-free structures which satisfy all the active criteria described above (*BLAminus*, ActLoopNT-in, SaltBr-in, intact APE5 helix). For the substrate-bound non-TYR kinase structures, the APE9(Cα)-hRd(O) distances range from 3.4 to 4.2 Å. This suggests that the Cα-O interaction is a CH-O hydrogen bond, which has been observed in proteins [41]. For 76% of human non-TYR catalytic kinases [42], the APE9 residue is a glycine, which forms Cα-O hydrogen bonds more readily than other amino acids likely for steric reasons. This contact is broken in many substrate-free structures, with the APE9(Cα)-hRd(O) distance exceeding 6 Å even those which satisfy other active-state criteria described above (**Figure 8A** bottom panels).

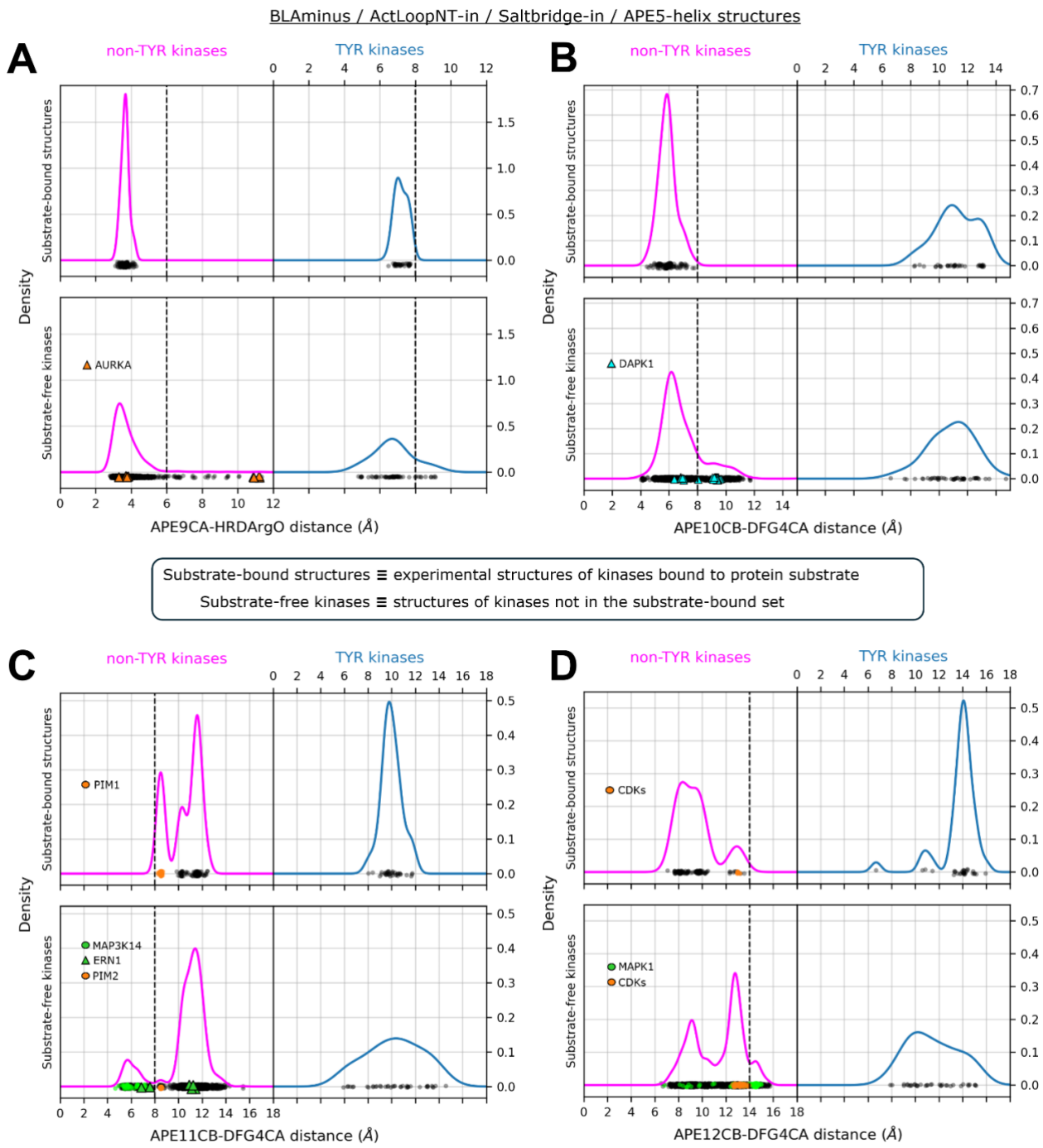


Figure 8. Distance criteria for Activation Loop C-terminal residues

Comparison of 248 substrate-bound structures vs substrate-free structures of the same kinases (as controls) which satisfy previously defined active-state criteria (*BLAminus*, SaltBr-in, ActLoopNT-in, APE5-helix intact) to identify the need for additional active-state criteria based on residue-residue distances that define the substrate binding surface.

A. Distribution of the APE9-Cα/HRD-Arg-O distance in substrate-bound structures (*top* panels) compared with substrate-free kinases (*bottom* panels). Structures of CAMK_AURKA are plotted with orange triangles. Inactive structures of AURKA greatly exceed the APE9-Cα/HRD-Arg-O distance threshold which was set at 6 Å.

B. Distribution of the APE10-Cβ/DFG4-Cα distance in substrate-bound structures (*top* panels) compared with substrate-free kinases (*bottom* panels). Structures of CAMK_DAPK1 are plotted with blue triangles. A dashed vertical line is drawn at the threshold of the APE10-Cβ/DFG4-Cα distance criteria, highlighting the presence of inactive DAPK1 structures (right of the dashed line).

C. Distribution of the APE11-Cβ/DFG4-Cα distance in substrate-bound structures (*top* panels) compared with substrate-free kinases (*bottom* panels). Structures of PIM family kinases (CAMK_PIM1 and CAMK_PIM2) are plotted with orange circles, structures of STE_MAP3K14 are plotted with green circles, and structures of OTHER_ERN1 are plotted with green triangles. A dashed vertical line is drawn at the lower threshold of the APE11-CB/DFG4-CA distance criteria (8 Å) to highlight the separation of active CAMK_PIM1/PIM2 (right of the dashed line) from inactive OTHER_ERN1 and STE_MAP3K14 (left of the dashed line).

D. Distribution of the APE12-Cβ/DFG4-Cα distance in substrate-bound structures (*top* panels) compared with substrate-free kinases (*bottom* panels). Structures of CDK family kinases (e.g., CMGC_CDK2) are plotted with orange circles and structures of CMGC_MAPK1 are plotted with green circles. A dashed vertical line is drawn at the upper threshold of the APE12-Cβ/DFG4-Cα distance criteria (14 Å) to highlight the separation of active CDKs (left of the dashed line) from inactive CMGC_MAPK1 (right of the dashed line).

Substrate-bound TYR kinase structures have somewhat longer APE9(Cα)-hRd(O) distances, ranging from 6 to 8 Å (**Figure 8A**, top right panel). This reflects a greater flexibility of the ActLoopCT for TYR kinases than for non-TYR kinases. Because APE9 interacts closely with the P+1 residue in TYR kinases, the upper bound on the APE9(Cα)-hRd(O) distance (8 Å) is likely important for allowing the substrate Tyr residue to reach the catalytic HRD-Asp residue. This is therefore the upper bound on this distance for TYR kinases.

APE10. Active TYR kinases differ significantly from non-TYR kinases in the APE9 – APE12 region to accommodate a specific binding mode for substrates [43]. TYR kinases typically have a bulky/charged residue at APE10 (most commonly Lys or Arg) with the side chain resting "on top" of the bound substrate peptide or protein, whereas for non-TYR kinases APE10 is placed "underneath" the substrate. In substrate-bound TYR kinases, the backbone of APE9 forms a short beta strand with P+1 of the substrate and APE10 and APE9 are both in the beta region of the Ramachandran map (**Figure 7B**). Hence, for active TYR kinases we require APE9 and APE10 (and APE8 described above) to have $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$. This is a broad region, but all active kinases identified by the full set of criteria are distributed across a narrower region (see below).

The APE8-APE10 dihedral criteria reveal the existence of inactive "ActLoopCT-out" structures of human TYR_SRC in the PDB: e.g., PDB: 1y57A and 1yi6B which are often considered "Active" in the literature but do not resemble substrate-bound structures of TYR kinases. The ActLoopCT conformation in these structures would likely interfere with substrate binding (**Figure 5E**). There is, however, a structure of chicken SRC (PDB 3dqw, chains A-C, **Figure 5E, top**) that is "ActLoopCT-in" and is quite similar to the ABL1 (PDB 2g2iB) structure listed in **Table 1**, and this structure is classified as "active" by our criteria. The chicken SRC sequence differs by only two amino acids in the kinase domain from human SRC, neither of which is in the activation loop. Hence, although there is no substrate-binding-capable structure of human SRC kinase domain in the PDB, the chicken SRC structure (PDB: 3dqW, chain A) presents a suitable active model of the human ortholog. We also note the existence of inactive ActLoopCT-out structures of TYR_JAK3-2 (JAK3 catalytic domain) in the PDB (e.g., 4rioA), which differ significantly from the substrate bound (autophosphorylation of Y836) structure of the closely related kinase TYR_JAK2-2 (6vnbA) (**Figure 5E, bottom**).

In non-TYR kinases, the side chain of APE10 is typically short and uncharged (most commonly

Val or Cys). In active and substrate-bound structures of non-TYR kinases, the APE10 and DFG4 side chains pack against each other to form a component of the substrate-binding surface that interacts with the substrate P-site and P+1 residue. To capture this structural motif in our criteria for active structures in non-TYR kinases, we use a cutoff of 8 Å on the Cβ-Cα distance between APE10 and DFG4 for non-TYR kinases alone (**Figure 8B**, upper left panel). This cutoff efficiently separates two populations of DAPK1 structures (cyan triangles in Figure 8B, lower left panel); the points below 8 Å resemble active structures of other non-TYR kinases, while those above 8 Å are substantially different and would block substrate binding. Changes in the conformation of APE10 from the substrate-compatible conformation often coincide with conformational changes of other residues as described below.

APE11 – APE12. For the majority of non-TYR kinases, activation loop phosphorylation (typically at APE12 or APE13, and less commonly APE11 [28]) stabilizes the N-terminal and C-terminal residues in a substrate-compatible conformation. Importantly, phosphorylation at these residues constrains the conformation of downstream residues in the ActLoopCT (APE12, APE11, APE10, APE9, APE8) which interact directly with substrates (**Figure 4**). For several non-TYR kinases, the requirement for phosphorylation on the activation loop as a requirement for catalytic activity can be explained by the unphosphorylated conformational propensities of the APE11 and/or APE12 residue(s), which may interfere with substrate binding. To capture this, we measured the Cβ-Cα distance of APE11-DFG4 and APE12-DFG4 and set bounds on the active criteria for non-TYR kinases based on the distribution of distances for substrate-bound structures (**Figure 8C and 8D**, upper left panel). These bounds are (8 Å, 14 Å) and (7 Å, 14 Å), respectively.

We find that, for several inactive non-TYR kinases in the unphosphorylated state, the Cβ vectors of APE10 and APE11 have "swapped" orientations resulting in an expanded distance between APE10 and DFG4 (Cβ – Cα > 8Å) and a contracted distance between APE11 and DFG4 (Cβ–Cα < 8Å), e.g., CAMK_DAPK1 (unphos. 3dgkA, phos. 2xuuA) (**Figure 5F)**, STE_MAP3K5 (unphos. 8qgyB, phos. 6e2mA), OTHER_ERN1 (unphos. 6w3bA, phos. 6w3cA) and CMGC_CDK2 (unphos. 1finA, phos. 1qmzA). Some structures violate only the lower bound of the APE11-DFG4 distance criteria, e.g., unphosphorylated CMGC_MAPK1 (4xj0A, 7w5oB) and STE_MAP3K14 (8yhwA). For comparison, CMGC_MAPK1 has phosphorylated structures in the PDB which resemble typical active structures of CMGC kinases (e.g., 2erkA). For STE_MAP3K14 however, there are no active structures of the same kinase to compare with because all existing PDB structures of MAP3K14 exhibit a contracted APE11-DFG4 distance (< 8 Å, similar to unphosphorylated ERN1 6w3bA) and have therefore been reclassified as "inactive" (**Figure 8C**, lower panel). Based on the observation that ERN1 and MAP3K14 have similar ActLoopCT sequences (^{APE12}Y**IPGTE**^{APE7} in MAP3K14 vs ^{APE12}G**VPGTE**^{APE7} in ERN1) we propose that the active state of MAP3K14 should resemble the structure of active ERN1 (6w3cA) (**Supplementary Figure 1**).

For CMGC_MAPK1 (ERK2), we observe an unphosphorylated activation loop conformation with

a distended C-terminus (PDB: 3erkA) that places the activation loop over the APE motif and in contact with the G-helix, as indicated by an expanded distance between APE12 and DFG4 (Cβ – Cα > 14Å) (**Figure 5F**). These structures, which are likely inactive, differ significantly from the phosphorylated state of MAPK1 (PDB: 2erkA) and other substrate-bound structures in the CMGC family, e.g., CDK2 (**Figure 8D**). The APE12 distance criterion proposed here designates 3erkA as inactive and 2erkA as active.

Finally, due to the flexibility of the activation loop C-terminus depicted for substrate-bound TYR kinases in **Figure 2** we forego the additional distance criteria for APE11 – APE12 for TYR kinases.

HRD Motif Conformation

98% of human catalytic protein kinases have a conserved Asp in the αF-helix that forms hydrogen bonds with the backbone of the HRD loop, which is in a strained conformation [44]. This structural motif stabilizes the positioning of the His and Arg residues, which themselves have a regulatory function. The side chain of HRD-His donates a hydrogen bond to the backbone carbonyl of X in the **X**DFG motif when it is in the *BLAminus* conformation [10], while HRD-Arg forms complementary charge-charge interactions with phosphorylated residues of the activation loop [13]. To include this structural motif in our criteria for structurally active kinases, based on visual inspection of the Ramachandran plots for the HRD motif over substrate-bound structures in the PDB (**Figure 7D**), we restrict active kinases to examples where the His-HRD backbone is located in the alpha region and the Arg-HRD backbone is located in the L region of the Ramachandran plot: $\phi \in (0°,180°)$, $\psi \in (-50°,100°)$. There is evidence from the PDB that this structural motif plays an important role in maintaining the active state: CAMK_AURKA has 64 inactive structures which are DFGinter, and the HRD Arg backbone nitrogen is flipped “out” in all these structures. This flip breaks hydrogen bonds with the αF-Asp side chain as in the DFGinter structure of AURKA, 2j50A (**Figure 5G**) [44]. Distortion of the HRD loop appears to be responsible for breaking the hydrogen bond between the HRD-His side chain and XDF carbonyl, allowing for a partial flip of the DFG motif from DFGin to DFGinter. Two structures of TYR_ABL1 (chains A and B of 2g2i), are “HRD-out”. The substrate has a Phe at the P0 position and bound ADP. While it is possible that the HRD-out state is similar to SaltBr-out in that it precludes catalytic activity but not the binding of substrate, it is notable that the only HRD-out substrate-bound structures are bound to a substrate-mimic (Phe) which fails to satisfy catalytic interactions with the HRD motif. This suggests that the HRD-in backbone conformation may be favored structurally by the interaction of HRD-Asp with substrate hydroxyl atoms.

The αF-Asp residue is an aspartic acid in every human catalytic kinase except for: CAMK_PIM1, CAMK_PIM2, CAMK_PIM3, OTHER_HASPIN, OTHER_PKDCC, TYR_AATK (LMTK1), TYR_LMTK2, and TYR_LMTK3. PIM1-3 have an Ala or Thr substituted at this position, meanwhile the other kinases in this list have Asn. From the PDB we observe that active structures of PIM1 and PIM2 contain a pair of water molecules which occupy the cavity formed by the substitution with Ala (as in PIM1, PDB: 2bzkB) or Thr (as in PIM2, PDB: 4x7qB). These water molecules appear to form hydrogen bonds with the HRD

backbone analogous to the hydrogen bonds formed by the carboxylate side chain of αF-Asp.In PDB structures of the PIM family kinases, the His and Arg backbone dihedrals are similar to all other active structures (A-L regions of the Ramachandran map). Nevertheless, we forego the HRD dihedral angle criteria for all kinases that do not have the conserved Asp at the αF helix.

Regulatory spine

Finally, we evaluated the utility of the regulatory spine for identifying active structures. The regulatory spine consists of four amino acids:

1) the His residue of the HRD motif. This residue is: His in 393 catalytic kinases; Tyr in 38 AGC kinases, CK1_CSNK1G1,2,3; OTHER_SBK2; TKL_LRRK2; Leu in OTHER_PKDCC, Phe in TKL_LRRK1.
2) the Phe residue of the DFG motif. This residue Phe except: Leu in 38 catalytic kinases; Tyr in 11 catalytic kinases; Trp in CMGC_CSNK2A1,2,3; Met in CMGC_CDK8,19; Val in OTHER_PBK.
3) the Glu4 residue (corresponding to CAMK_AURKA Q185), which is four positions after the conserved Glu of the N-terminal domain salt bridge. In catalytic kinases, this residue is: Leu (243 kinases); Met (114); Tyr (22); His (18); Phe (10); Ile (8); Gln (6); Cys (5); Val (4); Gly (3); Asn (2); Ser (2); Ala (2); Thr (1).
4) a usually hydrophobic residue just before the β4 strand, corresponding to L196 in CAMK_AURKA. We define this as "HPN7," which means the seventh residue from the start of the conserved HPN motif (HPNxxx***X***), which occurs in the loop between the C helix and the β4 strand. In catalytic kinases: Leu (256); Phe (58); Tyr (50); Met (25); Val (16); Ile (15); Cys (7); Ala (6); Thr (3); Gln (1); Ser (1).

These four residues define three distances: Spine1 (HRD-His/DFG-Phe), Spine2 (DFG-Phe/Glu4), and Spine3 (Glu4/HPN7). When the residues are small or polar, there may not be a contact between the side chains and such a contact may not be necessary for constructing an active kinase structure. In **Supplementary Figure 2**, the distribution of Spine1, Spine2, and Spine3 are shown for ATP-bound and unbound structures in the *BLAminus* and other states. From all three plots, it can be observed that nearly all ATP-bound, *BLAminus* structures contain an intact spine. The only exceptions are the Spine2 distances in two ATP-bound structures of PAK4 (PDB: 7s46, 7s47) in which the C-helix is twisted by about 45° starting at the residue before the salt-bridge glutamic acid (E366). This distorts the position of the M370 side chain, which forms the Spine2 distance with DFG-Phe's side chain. It is not known whether this distortion makes these PAK4 structures inactive, since the position of the Glu4 residue does not directly mediate contacts with ATP or the substrate.

Distributions of maximum spine distances (across Spine1, Spine2, Spine3) for kinase structures with and without ATP and in the *BLAminus* and other states is shown in **Supplementary Figure 2D**. As we described in our previous paper, a majority of structures in several DFGin conformational states

(*BLAminus*, *ABAminus*, *BLBplus*, etc.) contain an intact spine (defined as having all three spine distances less than 5.0 Å). 99% of *BLAminus* structures contain an intact spine. Most of those with a broken spine occur because of the position of the HPN7 residue, which does not interact with the substrate or ATP. When we apply the 5 active criteria described in the previous sections (DFGin, *BLAminus*, SaltBr-in, ActLoopNT-in, ActLoopCT-in), there are 2451 human catalytic kinase chains in the PDB that are "Active." There are only 14 structures in this set with a Spine3 distance greater than 6 Å, all of them PAK4 with a twisted end of the C-helix. There are only 11 active structures with a spine distance between 5 and 6 Å. For the sake of simplicity, we therefore do not use the regulatory spine as a criterion for active structures.

Active structures of catalytic kinases in the Protein Data Bank

From the considerations above, we define probable "Active" structures of kinases as those capable of binding ATP, Mg^{2+} ions, and substrate, with the criteria summarized in Table 2. From these criteria, we can label each feature required for the kinase to be "Active": 1) DFGin; 2) *BLAminus*; 3) SaltBr-in; 4) HRD-in; 5) ActLoopNT-in (DFG6 residue); 6) ActLoopCT-in (APE6-APE12 criteria).

We made certain exceptions to the criteria for some kinases. The salt bridge criterion is skipped for OTHER_WNK1, WNK2, WNK3, and WNK4 kinases (WNK- "With No Lysine") and for TKL_MAP3K12 and TKL_MAP3K13. In the experimental structures of TKL_MAP3K12 (e.g., 5CEP), the residue equivalent to the salt bridge Glu is Asp161 and is turned outwards with a break in the alpha C-helix, which is shorter than that of other kinases. The presence of the Asp makes the salt bridge less likely to form so we omitted it as a criterion for these two kinases. Finally, OTHER_HASPIN, OTHER_TP53RK, and OTHER_PKDCC do not have APE motifs [10], and do not fold into the same structures as the C-terminal regions of other kinases. Thus, there is no ActLoopCT requirement for these kinases. They are listed as noAPE in Table 2.

The six conformational labels (DFGin, *BLAminus*, SaltBr-in, ActLoopNT-in, ActLoopCT-in, HRD-in) are determined by a total of 26 structural features, enabling us to apply the criteria broadly to diverse families of kinases with different activation loop sequences. This set of features still allows for substantial structural variability within ensembles of experimental or simulated structures that represent the active form of protein kinases. This comes at the potential cost of including false positives in the active set, which we refer to as "false actives". To estimate the prevalence of false actives using our classifier, we calculated RMSD of the activation loop N-terminus and C-terminus (first 9 and last 15 residues: DFGXXXXXX and XXXXXXXXXXXXAPE) for active structures with respect to 32 kinases from **Table 1**, using one representative substrate-bound structure per-kinase. We focus on this subset of activation-loop residues as the frequency of interaction with substrates drops off significantly towards the middle region of the activation loop (**Figure 4**), and it is therefore unlikely that this outer segment of the activation loop defines kinase activity directly.

Table 2. Summary of criteria for classifying protein kinase domain structures as "Active"

Type *	DFGin-BLAminus	**KE Salt bridge	HRd ϕ,ψ‡	DFG6/ XHRD hbond	APE7,6 ϕ,ψ †	APE8 ϕ,ψ	APE9 ϕ,ψ	APE10 ϕ,ψ	APE8 χ_1	APE9/ xHRD Cα/O dist	APE10/ DFG4 Cβ/Cα dist	APE11/ DFG4 Cβ/Cα dist	APE12/ DFG4 Cβ/Cα dist
non-TYR	X	<3.6 Å	AL	<3.6 Å	AA/BL	B	-	-	g^-	<6 Å	<8 Å	8-14 Å	7-14 Å
TYR	X	<3.6 Å	AL	<3.6 Å	AA/BL	B	B	B	-	<8 Å	-	-	-
noAPE	X	<3.6 Å	AL	<3.6 Å	-	-	-	-	-	-	-	-	-

* TYR are all TYR family kinases, whether they phosphorylate Tyr or Ser/Thr or both. non-TYR are all other protein kinases, whether they phosphorylate Tyr or Ser/Thr or both. noAPE consists of the kinases without the APE motif and associated helix (OTHER_PRKDCC, OTHER_TP53RK, OTHER_HASPIN).

** Only applied if the residues in these conserved positions are actually Lys/Arg and Glu (e.g. WNK kinases and MAP3K12/13 are exempt).

†For **ϕ,ψ** in the table, the letters correspond to broad regions of the Ramachandran map

A = $\phi \in (-180°,0°)$, $\psi \in (-100°,50°)$
B = $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$
L = $\phi \in (0°,180°)$, $\psi \in (-50°,100°)$
g^- = $\chi 1 \in (-120°,0*)$

‡ "AL" means His is in A region and Arg is in L region of Ramachandran map respectively: only for kinases that have Asp in F-helix (D in motif D-[ILVMW]-[WF]).

We find that only one of the kinases from **Table 1**, CMGC_CDK2, has X-ray or cryo-EM structures in the PDB which are active by our criteria and deviate from the substrate-bound structure by more than 2 Å in the N-terminal and C-terminal regions of the activation loop. These outliers comprise only 6% of the structures classified as active for this kinase: 2c4gC, 2bpmC, 2wihC, 1vywC, 2wpaC, 2bkzC, 2bkzA, 3eocA, and belong to a common active-like, unphosphorylated conformation of CMGC_CDK2 (**Supplementary Figure 3**). For the remaining 31 kinases from **Table 1** with fully resolved N-terminal and C-terminal activation loop residues, structures predicted to be active always fall within 2 Å RMSD of the substrate-bound state.

Table 3. Classification of human catalytic kinase domain structures in the PDB

	With or without ActLoop disorder			With no ActLoop disorder		
	Chains	Catalytic human kinase domains	Percent (of 437)	Chains	Catalytic human kinase domains	Percent (of 437)
Any conformational state	9881	277	63.4%	5547	224	51.3%
DFG-in	8490	261	59.7%	5075	207	47.4%
DFG-in+BLAminus	5401	214	49.0%	3894	170	38.9%
DFG-in+BLAminus+HRD-in	5076	205	46.9%	3589	161	36.8%
DFG-in+BLAminus+SaltBr-in	4338	195	44.6%	3282	154	35.2%
DFG-in+BLAminus+ActLoopNT-in	5224	206	47.1%	3821	167	38.2%
DFG-in+BLAminus+APE6-APE7	4902	192	43.9%	3606	155	35.5%
DFG-in+BLAminus+APE6-APE8	3891	168	38.4%	2963	142	33.6%
DFG-in+BLAminus+APE6-APE9	3658	156	35.7%	2893	136	31.1%
DFG-in+BLAminus+APE6-APE10	2555	118	27.0%	2170	106	24.3%
DFG-in+BLAminus+APE6-APE11	2521	114	26.1%	2161	105	24.1%
DFG-in+BLAminus+APE6-APE12	2369	112	25.6%	2028	103	23.6%

Active	2402	130	29.8%	2180	116	26.5%

Counts obtained for human catalytic kinases in the PDB as of January 21, 2026

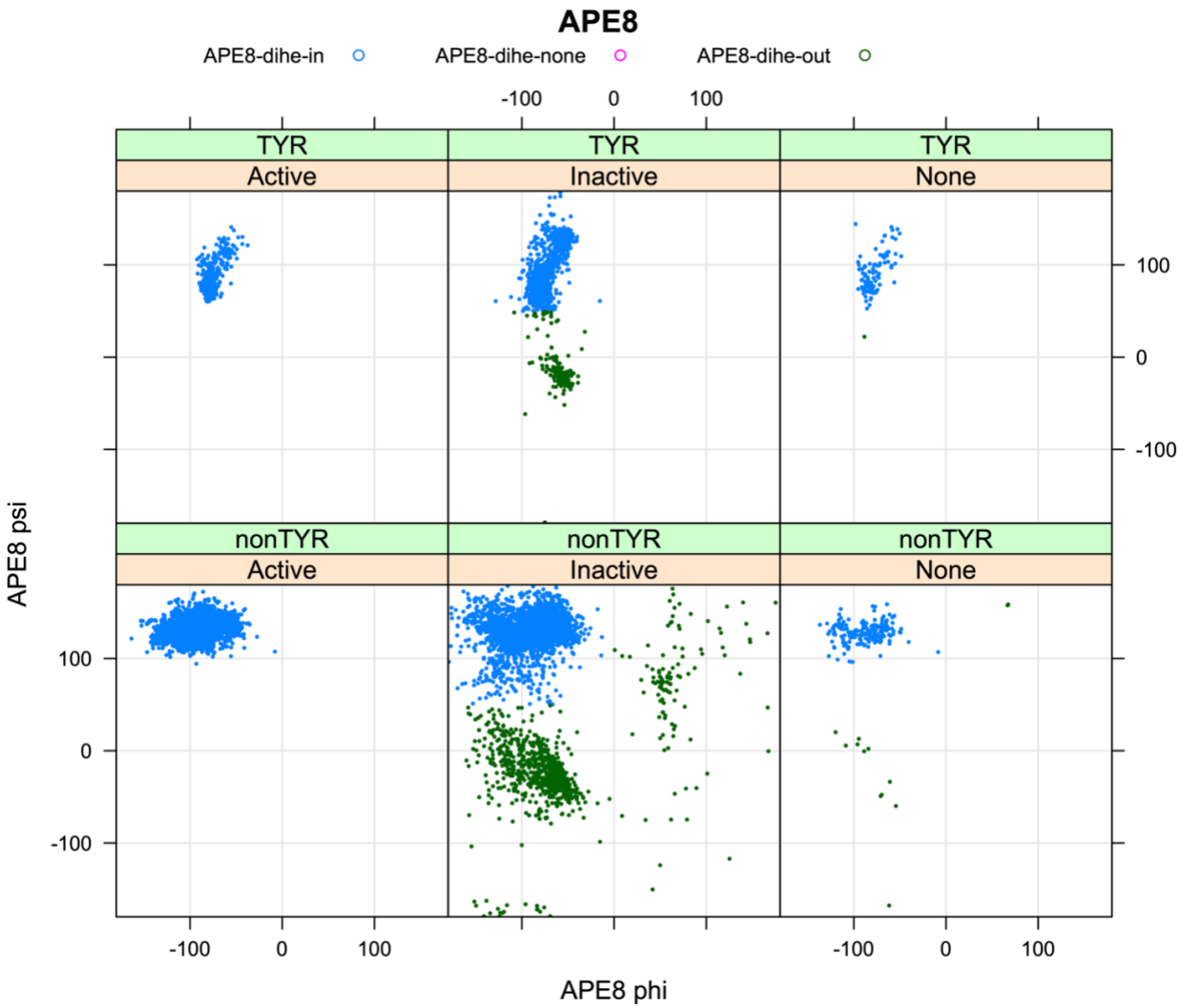


Figure 9. Ramachandran plots for residue APE8 of 10,166 human protein kinase domains in the PDB.

The structures are divided into the TYR (top row) and non-TYR groups (bottom row) and the designation of the kinase domain as Active, Inactive, and None (missing atoms). Blue points are considered APE8-dihe-in and green points are APE8-dihe-out. Even though the criteria for APE8 is a broadly defined B region ($\phi \in (-180°,0°)$, $\psi \in (50°,180°)$), the distribution in active kinases is much more limited (compare distribution of blue dots in Active vs Inactive kinases).

Using the criteria listed in Table 2, we labeled all human protein kinase domain structures in the PDB as “Active”, “Inactive”, or “None” to indicate missing data (e.g., missing residues). We labeled the structure as “Inactive” if any of the criteria were violated. “Active” structures must be DFGin-*BLAminus*, HRD-in, and Salt-bridge intact (or at least Chelix-in), since structures with missing atoms in these motifs cannot easily be used for further analysis (such as MD simulations). If ActLoopNT or ActLoopCT was labeled “out”, then the structure was labeled Inactive. If none of the criteria were labeled “out” but the ActLoopNT or ActLoopCT were missing atoms, then the structure was labeled “None.” The results are shown in **Table 3** for the catalytic kinases. Of the 437 catalytic kinase domains in the human proteome, only 130 (29.8%) have active structures in the PDB. Of these, only 116 have complete sets of coordinates

for the backbone of the activation loop, comprising less than 30% of catalytic kinases in the human proteome.

Finally, for all the 10,166 labeled human protein domains in the PDB, we examined the dihedral angle distributions for the HRD residues and the APE6-APE10 residues, plotting Ramachandran maps for each residue given the activity state of the kinase (Active, Inactive, None), separately for TYR and non-TYR kinases (**Figure 9, Supplementary Figure 4**). Even though we used broad regions of the Ramachandran map for the HRD residues and the APE6-APE10 residues, the distributions are much narrower for the Active kinases, indicating that the various criteria are not completely independent. For example, for residue APE8, the criterion is that the backbone dihedral angles are restricted to a broad Beta region of the Ramachandran map: $\phi \in (-180°,0°)$, $\psi \in (50°,180°)$. But for the active kinases, the ϕ,ψ values are much more narrowly distributed (**Figure 9**) with $\psi>100°$ for almost all non-TYR kinases and $\phi>-100°$ for TYR kinases.

Generation of active models of catalytic protein kinase domains

To generate active models of the 437 human protein kinase domains, we created sequence sets for the multiple sequence alignments (MSAs) required by AlphaFold2 and template data sets in the active form. Sets of orthologous sequences (or near paralogs) for each kinase were created from UniProt such that each sequence in an ortholog set for a given kinase was greater than 50% identical to the target and aligns to at least 90% of the target kinase domain length with fewer than 10% gaps. Each ortholog set was filtered with CD-HIT [45] so that no two sequences in each set were more than 90% identical to each other. This was done to create diversity within the ortholog sets for each kinase. We also created “Family” sequence sets consisting of all the human kinase domains within each kinase family.

To create a set of templates, we identified all active structures of catalytic kinases in the PDB as of August 2024 (including non-human kinases) using the criteria given above and, for each kinase, selected a structure with an activation loop similar to one of the 248 substrate-bound structures (< 1.5 Å RMSD of the first five and last 15 residues of the activation loop). If more than one similar structure was available, we chose the template structure with the largest number of coordinates for the activation loop residues (to select in favor of complete activation loops). When possible, we chose a template structure which is bound to ATP or an ATP analog. This resulted in a set we named “ActivePDB”, consisting of 101 kinase domains (one structure per-kinase).

We used the Colabfold implementation of AlphaFold2 [46] to model 437 human catalytic kinase domains in the active conformation, using the ortholog and family sequence sets and each individual template from the ActivePDB set. Self-templates (where the template kinase is the same as the query kinase) were skipped to accurately assess the performance of our method. For each run of Colabfold (1 seed per run) we utilized AlphaFold2 Multimer (v3) weights. One benefit of using AlphaFold-Multimer is that all 5 models utilize provided templates (in original AlphaFold2, only models 1 and 2 use templates

as input features, models 1.1.1 and 1.1.2 in Jumper et al. [18]). After each run, the models were relaxed with AMBER and the standard AlphaFold2 protocol. As a baseline, for each run of Colabfold we used a maximum MSA depth of 50 sequences. This approach yielded active models for 426 kinases in aggregate: 417 kinases using ortholog MSAs, and 415 kinases using family MSAs. The remaining 18 kinases which were resistant to adopting the active conformation at the baseline MSA depth of 50 sequences required smaller depths (25 and 10 sequences) to generate active models. Several kinases were sufficiently resistant to the active templates as to require drastically smaller MSA depths (2 or 3 sequences): TYR_AATK (LMTK1), TYR_LMTK2, TYR_LMTK3, OTHER_WEE1, OTHER_WEE2, OTHER_PKDCC, and CAMK_OBSCN-2. The kinase OBSCN-2 required additional treatment as described below.

Table 4. Number of active catalytic kinase domains (out of 437) produced by different Template and Sequence data sources

Template Sources	Sequence Sources	Number of active kinases
EBI (PDB70)		*240*
None	Uniref30	361
ActivePDB	Family	430
ActivePDB	Ortholog	435
ActivePDB	Family+Ortholog	436
All	All	437

Table 4 summarizes the results for various template and MSA sequence sources. The results of our single-template protocol (using ortholog and family MSAs) yielded active models for all but one of the 437 catalytic kinases. The only kinase for which we could not generate active models using the single-template method, AGC_MASTL, contains a ~500 residue-long insertion in the activation loop (the longest activation loop in the kinome). For AGC_MASTL, we utilized Colabfold without the use of templates and with the MSA provided by the MMseqs2 server [47] and 101 seeds (the same number of models built with templates). This produced an Active model of MASTL.

For the second kinase domain of obscurin (CAMK_OBSCN-2), the C-terminal segment of the activation loop made an α helix of residues 7825-7829 in all models that blocked access to the substrate binding site. We made the mutation D7929G (residue APE9, which is conserved as Gly in 73 out of 83 catalytic CAMK kinases) which helped to unfold this helix, at an MSA depth of 2 sequences. It is possible that OBSCN-2 is a pseudokinase.

For LMTK family kinases (TYR_AATK, TYR_LMTK2, and TYR_LMTK3), all AlphaFold2 models produced from moderate to large MSA depths (50 or more sequences) formed a folded activation loop containing a strand-turn-strand motif that would be inconsistent with substrate binding. This structural motif exists in many DFGout structures of TYR kinase family members. Active models could be produced for these kinases only at significantly low MSA depths (2 or 3 sequences). We also note that LMTK2 has an asparagine residue in place of the C-helix glutamic acid residue of the N-terminal domain salt bridge. However, this kinase retains catalytic activity and has been shown to phosphorylate CFTR and other

substrates involved in neuronal activity [48].

To assess the added value of our structure prediction pipeline compared with what is already available to the public, we downloaded structures of all 429 kinases from the EBI database of AlphaFold2 models (all v6 models, containing 437 catalytic domains). The EBI database contains only one model per protein, and only 240 of the 437 kinase domains (55%) contain active structures within this set according to our criteria. By comparison, when running all five models within Colabfold using default parameters, MSAs from the MMseqs2 server [47], 101 seeds, and without templates, using Uniref30 as the sequence database we obtained active models of 361 catalytic kinases (82.6%).

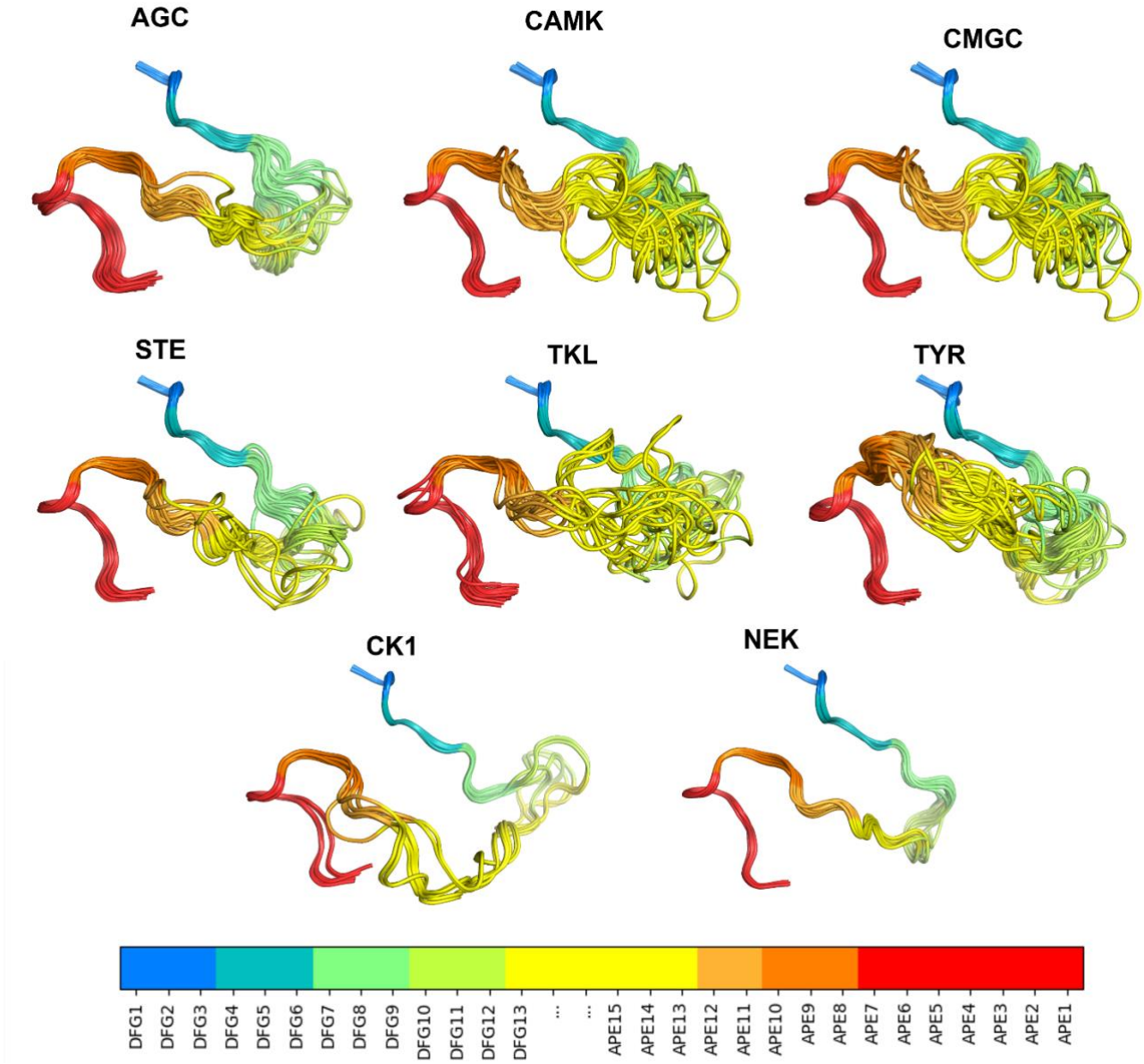


Figure 10. AlphaFold2 models of the active forms from each kinase family.
For clarity, some structures with long disordered regions within the activation loop are not shown in each family. Diverse kinases from the group labeled "OTHER" were also excluded from the visualization. In the legend (bottom), N-terminal activation loop residues are labeled DFG*n* where *n* is the nth residue from the beginning of the loop, starting from DFG Asp (*n*=1) and ending at DFG13 (*n*=13). C-terminal activation loop residues are labeled APE*n* in reverse order where *n* is the nth residue from the end of the loop, starting from APE15 (*n*=15) and ending at APE1 Glu (*n*=1).

Identifying the best model with interaction prediction scores of the activation loop

For each of the 437 kinases we sought to identify the active model generated by AlphaFold2 which has the most confidently predicted activation loop. We first defined an interaction confidence score for the activation loop using ipSAE (interaction prediction Score of Aligned Errors)[22]. ipSAE is a modified version of the AlphaFold ipTM score that depends on the PAEs (predicted Aligned Errors) between two sets of residues, e.g., two different proteins in a complex. Here we apply ipSAE intramolecularly to score activation loop residues j aligned on non-activation loop residues i in the kinase domain with pLDDT ≥ 60:

$$ipSAE_{Actloop} = \max_{\substack{i \notin \mathrm{ActLoop} \\ plDDT_i > 60}} \left[\underset{j \in ActLoop}{\mathrm{mean}} \left(\frac{1}{1 + \left(\frac{\mathrm{PAE}_{ij}}{d_0} \right)^2} \right) \right]. \quad (1)$$

Where PAE has units of Angstroms and d_0 is a scaling factor which, in typical ipSAE calculations between two interacting proteins, is a function of the length of the scored protein. Since the length of the activation loop varies from kinase-to-kinase and we wish to compare interaction scores on the same scale, we set d_0 to a fixed value of 4 Å which maximizes the variance of the scores of the predicted structures (**Supplementary Figure 5**).

The experimentally determined structures of substrate-bound kinases (**Figure 2**) show that, in active kinases, the activation loop is generally situated against the kinase domain and extends from the DFG motif towards the right-edge of the kinase domain (as usually oriented and shown in **Figure 2**). It then turns around and moves leftward and concludes in the APE motif, roughly below the DFG motif. This open U shape is characteristic of substrate-bound structures and of active AlphaFold2 models produced by our pipeline. The top-scoring AlphaFold2 models produced in this study are shown in **Figure 10** for each kinase family.

To benchmark the behavior of our pipeline in modeling active structures of catalytic kinases, we first compared the collection of AlphaFold2 models for the 29 kinases listed in **Table 1** that have complete activation loops within one of their experimental substrate-bound structures. When the same kinase is listed more than once in **Table 1**, we picked a single example since the activation loop structures were all very similar (<0.5 Å RMSD except for the two autophosphorylation structures of TKL_IRAK4). These experimental structures contain substrates and likely belong to an ensemble of substrate-binding-capable conformations of the activation loop of each kinase. Both experimental and computed structures that pass our "Active" tests still exhibit some heterogeneity of the structure of the activation loop, especially for residues far from the beginning or the end of the loop. This may be natural structural variation, and it

is possible or even likely that multiple conformations are compatible with substrate phosphorylation. In any case, we explored the ability of ipSAE (Eq. 1) to pick out good models that pass our "Active" tests described above. In **Figure 11**, we show scatterplots of RMSD vs ipSAE for these 29 kinases. The results demonstrate that for most of the kinases, the model with the highest ipSAE of the activation loop also tends to correspond to the lowest or very close to the lowest RMSD to the structures listed in **Table 1**.

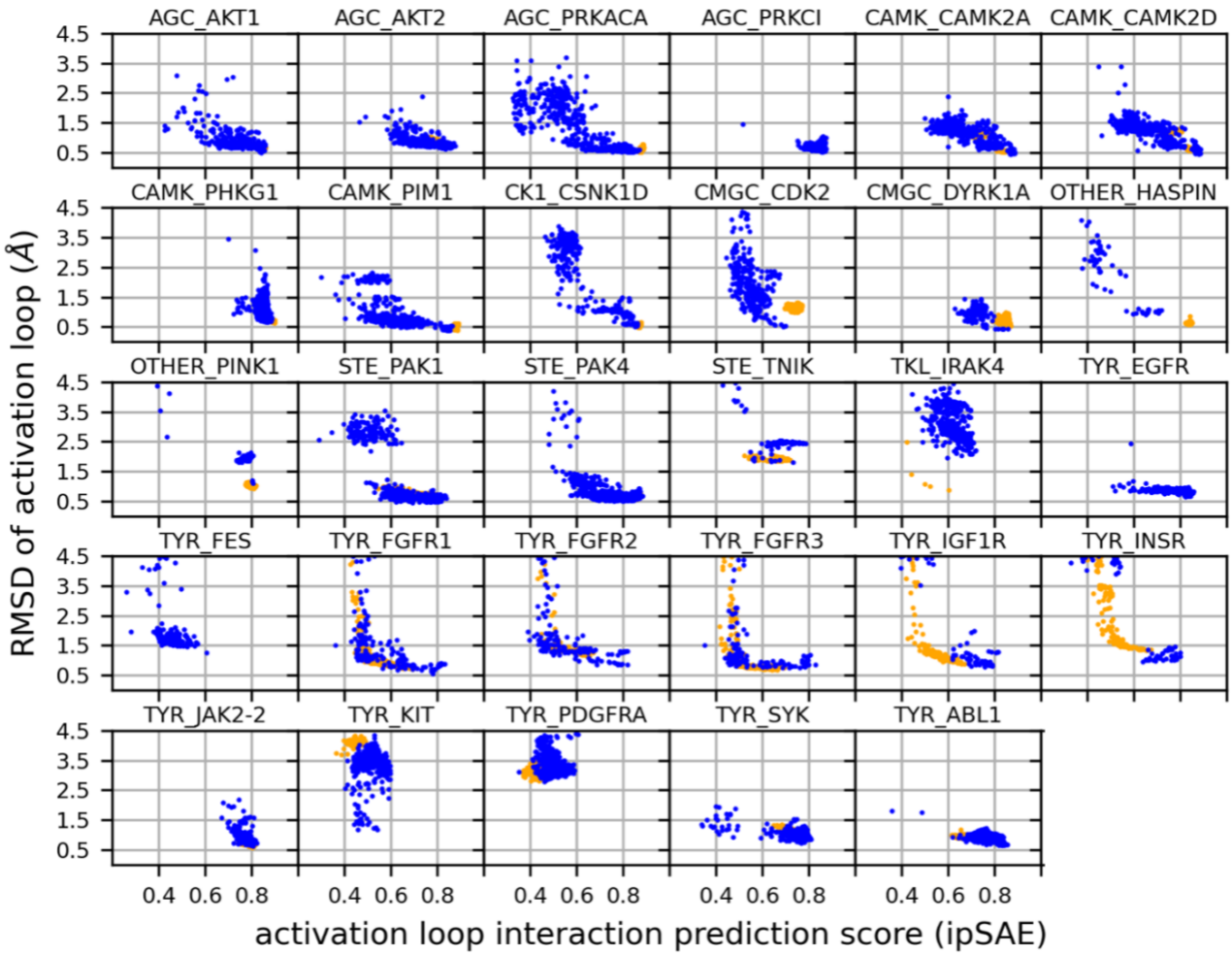


Figure 11. RMSD values for 29 substrate-bound structures from Table 1 versus activation loop interaction prediction score of each active AlphaFold2 model.

Models from AF2 with different template modes are shown in different colors: active structures from the PDB ("ActivePDB", blue), no templates provided to AF2 (orange). Models above 4.5 Å RMSD and intramolecular ActLoop ipSAE < 0.2 are not shown.

To extend the benchmark, we picked out at least one structure for each kinase in the August 2024 PDB with complete (or mostly complete) activation loops and extended this to orthologous kinases, resulting in 117 benchmark kinases (see Methods). The resulting benchmark set is comprised of the substrate-bound active structure (if available) or an active structure with the lowest activation loop RMSD to one of the substrate-bound structures. The distribution of RMSD to the highest scoring models (highest ipSAE score) and the distribution of ipSAE scores for these models are shown in **Figure 12**. The results show that 83 (71%) of the 117 kinases are represented by a model with less than 1 Å backbone atom

RMSD (N,Cα,C,O) over the whole activation loop, after superposition of the C-terminal domains of each kinase (**Figure 12A**). A total of 108 (92%) are less than 2.0 Å.

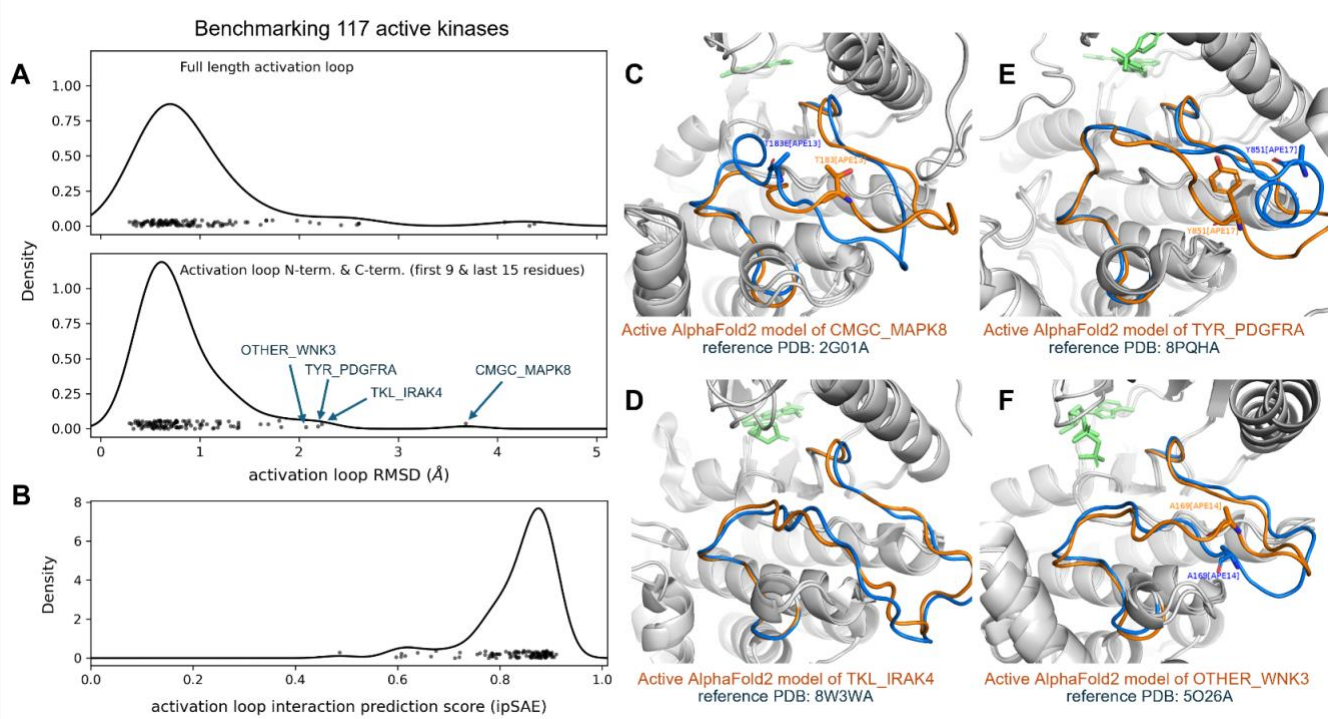


Figure 12. Benchmark of 117 active kinases in the PDB with full coordinates for the activation loop.

A. Distribution of RMSD values (with kernel density estimate) for the top scoring model, (top) RMSD of the full-length activation loop, (bottom) RMSD of the first 9 N-terminal residues (DFG – DFG9) and last 15 C-terminal residues (APE15 – APE) of the activation loop.

B. Distribution of activation loop confidence scores for these 117 AlphaFold2 models.

C-F. Shown examples of cases where the activation loop N-term/C-term RMSD exceeds 2 Å. Activation loops of active AlphaFold2 models are shown in orange, and benchmark PDB structures are shown in blue.

C. The CMGC_MAPK8 benchmark structure 2g01A deviates significantly from the active AlphaFold2 model, likely due to mutations made in the structure (T183E[APE13]).

D. The AlphaFold2 modeled activation loop of TKL_IRAK4 exhibits small, evenly distributed deviations from the benchmark structure 8W3WA.

E. The benchmark structure for TYR_PDGFRA, 8pqhA, is substrate-bound but unphosphorylated at Y841[APE17] which allows a short helix to form in the middle of the activation loop. The AlphaFold2 model, on the other hand, appears to resemble the phosphorylated state of closely related TYR kinases, e.g., FGFR2 (3gqiA) phosphorylated on Y654. In the AlphaFold2 model of PDGFRA, the equivalent residue Y841 points towards the HRD motif.

F. OTHER_WNK3 exhibits small yet distributed deviations in the activation loop modeled by AlphaFold2, compared with the benchmark structure 5o26A.

Flexibility of the middle-region of the activation loop (between DFG9 and APE15) is expected. Conformational variability in this region is likely a natural consequence of the ensemble nature of the active state, as the frequency of interaction between activation loop and substrate drops off significantly between DFG5 and APE13 (**Figure 4**). Therefore, to conduct a more appropriate assessment of the limitations of our structure prediction pipeline we analyzed the RMSD of just the first 9 and last 15 residues of the activation loop (DFGXXXXXX and XXXXXXXXXXXXXAPE), which are likely the most

structurally constrained features of the active ensemble. We refer to this measure of RMSD as "NT/CT RMSD". The results displayed in the bottom panel of **Figure 12A** show that, for 89 of the 117 benchmark kinases (76%) are within 1.0 Å of an experimental structure, and 113 of the 117 benchmark kinases (97%) our AlphaFold2 models are within 2.0 Å RMSD. The active kinase models with > 2.0 Å RMSD are: CMGC_MAPK8, TKL_IRAK4, TYR_PDGFRA, and OTHER_WNK3 (**Figure 12C-F**).

To further explore the limitations of our active criteria and scrutinize our AlphaFold2 models, we investigated each of the benchmarking outliers from **Figure 12A** as well as other examples where the AlphaFold2 model deviates from one or more active structures in the PDB by > 2 Å (ActLoop NT/CT RMSD). In total, there are 12 kinases which have at-least one structure in the PDB other than the benchmark that pass our active criteria but deviate from the corresponding predicted active AlphaFold2 model by more than 2 Å RMSD in the first 9 and last 15 residues of the activation loop. Each discrepancy between the predicted model and PDB structure may fall into one of the following cases:

Case A: The PDB structure is actually inactive, the AlphaFold2 model is active.
Case B: The AlphaFold2 model is actually inactive, the PDB structure is active.
Case C: Both the PDB structure and the AlphaFold2 model are active (different active states or variations within the active ensemble)
Case D: Both the PDB structure and AlphaFold2 model are actually inactive (different inactive states)

An overview of the 12 kinases with larger discrepancies to at least one active PDB structures is given in **Supplementary Table 5,** and the structures are shown in **Supplementary Figure 6**.

Case A occurs when the PDB structure is labeled active by our criteria but is likely not compatible with substrate or catalysis. Five discrepancies were assigned to **Case A**: CMGC_MAPK8, AGC_PDPK1, CMGC_CDK2, CMGC_MAPK1, and CK1_CSNK1D. For each of these discrepancies, the AlphaFold2 model resembles the active benchmark and/or substrate-bound structure, and structural deviations for one or more "active" PDB structures occur in critical substrate-binding motifs. The RMSD of our active CMGC_MAPK8 model to its benchmark PDB structure 2g01A is the most significant outlier, and upon closer examination of the structures we argue that the "active" benchmark structure of CMGC_MAPK8 (2g01A) is actually inactive. 20g1A is mutated to a phosphomimetic at the Thr[APE13] phosphorylation site (T183E) which contorts the C-terminus of the activation loop (APE13 through APE9). The activation loop of this structure has an NT/CT RMSD of 4.6 Å with respect to the active (and phosphorylated) benchmark of a closely related kinase, CMGC_MAPK1 (5v60A). In contrast, the NT/CT RMSD of 5v60A to our AlphaFold2 model of CMGC_MAPK8 is 1.9 Å, suggesting our model of CMGC_MAPK8 is likely a better representation of the active state than 2g01A. There are no wildtype structures of CMGC_MAPK8 in the PDB which pass our active criteria – all structures of this kinase that pass our active criteria are T183E mutants, and all have RMSDs greater than 2 Å from the active AlphaFold2 model. We argue that these "active-like" PDB structures of CMGC_MAPK8 are a result of noise at the boundary of our APE9

criteria: the average APE9-Cα/HRD2-O distance in these structures is 4.9 Å – an unusually large value for substrate-bound structures of CMGC kinases (which is at most 3.3 Å: see **Table 1**).

Only one discrepancy, TKL_BMPR1B, was assigned to **Case B**: in this case the active PDB structure of TKL_BMPR1B, 3mdyC, resembles the benchmark active structure of a closely related kinase (TKL_TGFBR1: 5friA) while the AlphaFold2 model of TKL_BMPR1B deviates significantly at APE10 (**Supplementary Figure 7**), which is an important substrate-binding site. This model was generated using CAMK_DAPK3 as a template. In the online dataset on the Kincore website, we replaced this model with the next highest-scoring AlphaFold2 structure (ipSAE of 0.771 to 0.769) which instead uses TKL_IRAK4 as a template, resulting in an orientation of APE10 that is more consistent with the active TGFBR1 PDB structure (**Supplementary Figure 7**).

Six discrepancies were assigned to **Case C**: OTHER_WNK3, CMGC_CLK1, CMGC_DYRK1A, TKL_IRAK4, TKL_TGFBR1, and TYR_PDGFRA, because the active AlphaFold2 model *and* PDB structures that deviate from this model by more than 2 Å both appear to be active. Structural deviations for these cases occur towards the middle of the activation loop, which are unlikely to interfere with substrate binding. We have not assigned any discrepancies to **Case D**. In summary, we identified 12 kinases with at least one structure that is labeled active but deviates from the corresponding AlphaFold2 model by more than 2 Å RMSD. Of these 12 discrepancies, 5 appear to be examples of **Case A**, 1 of **Case B**, and 6 of **Case C** (**Supplementary Table 5**).

For some other kinases, the AF2 models have poor ipSAE scores for the activation loop. This occurs for some kinases that are remotely related to other kinases in the human proteome or that have particularly long activation loops. The following 20 kinases have a top-scoring model with ipSAE < 0.5, in order of increasing score: AGC_MASTL, OTHER_CDC7, OTHER_PKDCC, AGC_LATS1, AGC_LATS2, TYR_LMTK2, TYR_NTRK3, OTHER_WEE1, OTHER_WEE2, AGC_STK38L, TYR_AATK (LMTK1), TKL_RIPK1, TYR_LMTK3, OTHER_EIF2AK4-2, AGC_STK38, TKL_LIMK2, TYR_MST1R, TYR_FRK, TKL_LIMK1, and OTHER_EIF2AK3. We note that OTHER_CDC7 bears close resemblance to the substrate-mimetic complex in **Table 1** (0.5 Å NT/CT RMSD with PDB: 6ya7A) despite its low score of 0.25.

There are only two examples of catalytic kinases which bear an unusual hydrophobic residue at the HRD-His position: TKL_LRRK1 substitutes His with Phe, and OTHER_PKDCC substitutes His with Leu. While the Phe sidechain of TKL_LRRK1 lacks hydrogen bonding capabilities, our AlphaFold2 models predict an interaction between the X[XDFG] backbone carbonyl (in the *BLAminus* state) and the "tip" of the Phe ring (Cζ atom). On the other hand, OTHER_PKDCC has a Leu sidechain at the His-HRD position which is too short to interact with and stabilize the XDFG backbone in the *BLAminus* conformation according to our AlphaFold2 models. Without this stabilization of *BLAminus* provided by the HRD loop, AlphaFold tends to generate structures in the *ABAminus* conformation even when provided *BLAminus* structures as templates. In this conformation, the activation loop of OTHER_PKDCC C-

terminus forms a beta sheet that folds against the αF-helix (**Supplementary Figure 8A**).

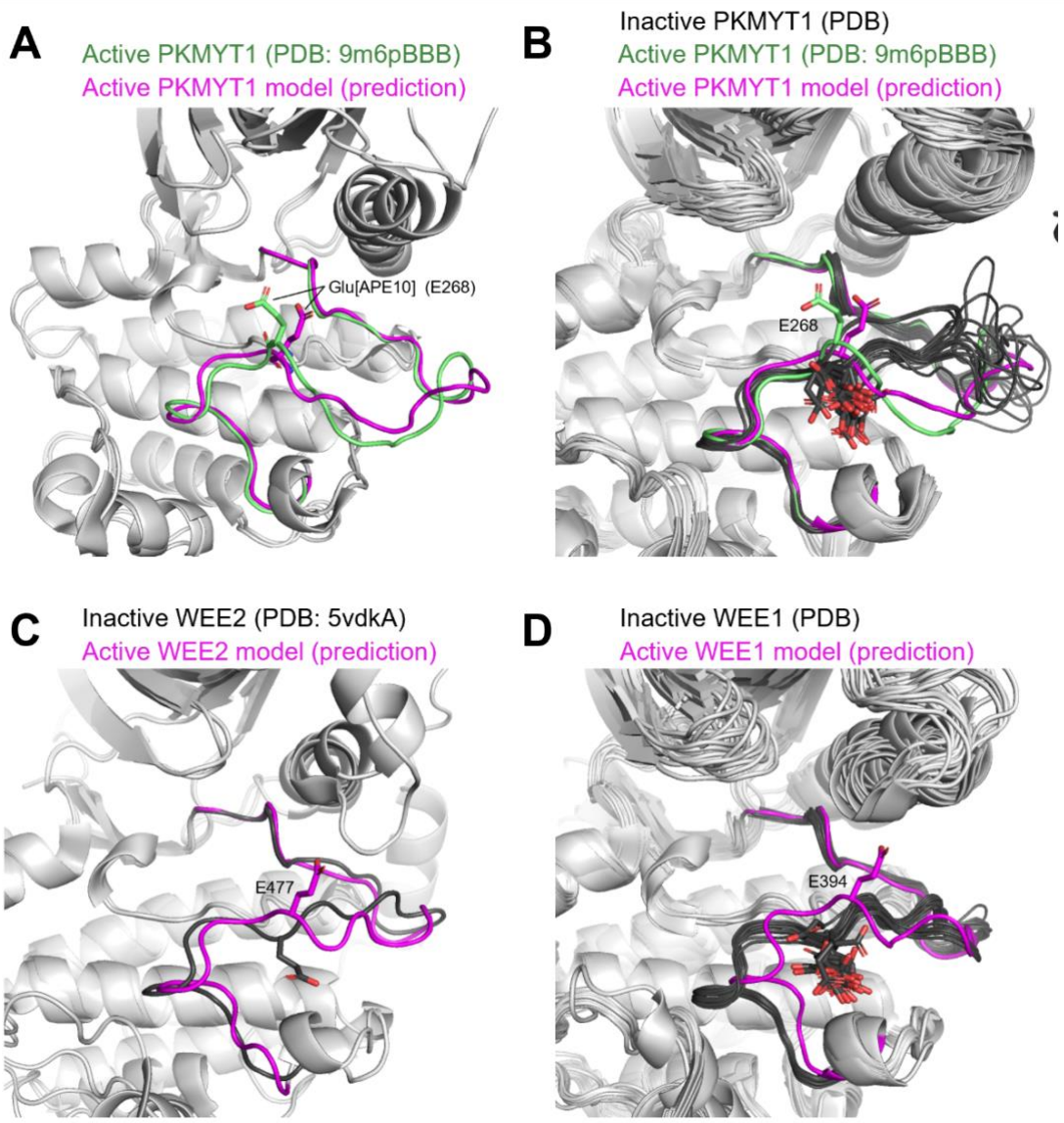


Figure 13. Validating the predicted active conformation of PKMYT1, WEE1 and WEE2

A. Superposition of a recently experimentally solved structure of OTHER_PKMYT1 (PDB: 9m6pBBB, *green*) with our predicted active model of this kinase generated with AlphaFold2 (*magenta*). The first 9 and last 15 residues of the activation loop have an RMSD of 1.3 Å. Glu[APE10] (E268 in PKMYT1) is shown with sticks; both structures are in an active "APE10-in" conformation typical of substrate-bound non-TYR kinases.

B. All PDB structures of PKMYT1 superimposed onto our active model for this kinase (*magenta*). The recently experimentally solved active structure of PKMYT1 is shown in *green*. All other experimental PDB structures of this kinase have Glu[APE10] (E268) flipped "out" and are classified inactive/APE10-out by our criteria (*black*). The active PDB structure of PKMYT1 (*green*) validates our predicted active structure (*magenta*) and supports the classification of other PKMYT1 structures as inactive.

C. All PDB structures of WEE2 (*black*) superimposed with our active model of this kinase which was generated with AF2 (*magenta*). All PDB structures of this kinase have Glu[APE10] (E394) flipped "out" and are classified as inactive. There is currently no experimentally determined active/APE10-in structure of WEE2 in the PDB according to our criteria.

D. All PDB structures of WEE1 (*black*) superimposed with our active model of this kinase which was generated with AF2 (*magenta*). All PDB structures of this kinase have Glu[APE10] (E477) flipped "out" and are classified as inactive. There is currently no experimentally determined active/APE10-in structure of WEE1 in the PDB according to our criteria.

Only one *BLAminus* structure was generated by our structure prediction pipeline for this kinase,

and in this case the activation loop scores very poorly. We hypothesize that the former case, where OTHER_PKDCC is in the *ABAminus* state with the activation loop folded against the αF-helix, may indeed be the active state for this kinase (**Supplementary Figure 8B**). This may be the only exception in the catalytic human kinome where the *ABAminus* conformation can be regarded as the most likely active conformation. In support of this prediction, using AlphaFold3 to generate structures of OTHER_PKDCC bound to ATP and Mg^{2+} we observe the same conformation of the activation loop and *ABAminus* state of the XDFG residues (**Supplementary Figure 8C-D**).

Post hoc benchmarking and predicted active structure for Wee1 family kinases

Since the completion of this work, we have identified experimental structures for two new kinases that were released by the PDB in January 2026 and are classified as active using our structural criteria: OTHER_PKMYT1 (9m6pBBB) and CMGC_MAPK14 (9cj2A). These kinases did not have any experimentally determined active structures (according to our criteria) at the time of our structural bioinformatics analysis on the August 2024 PDB and modeling computations and can therefore be used to test our structural predictions *post hoc*. Our active models of these kinases closely resemble their recently solved active structures (RMSD ≈ 1.3 Å for the first 9 and last 15 residues of the activation loop).

OTHER_PKYMT1 belongs to a functionally distinct family of kinases that includes OTHER_WEE1 and OTHER_WEE2. Notably, there are no substrate-bound structures of these kinases, and they are somewhat distantly related to any structure in the ActivePDB template set. These kinases are unique in that they more closely resemble non-TYR kinases in both sequence and structure but paradoxically exhibit highly specific Tyr kinase activity (Y15 in CDK1) [49], presenting an interesting test case for our classification method outside the typical TYR vs non-TYR kinase paradigm. It is not known whether this family binds substrate like typical non-TYR families (most Ser/Thr kinases) or TYR kinases as depicted in **Figure 2**.

Due to their closer sequence homology to non-TYR kinases, for this family we apply our active criteria conditioned on substrate-bound non-TYR specific kinases. Our active model for PKMYT1 adopts the same conformation as the recent experimentally determined PKMYT1 structure, PDB: 9m6pBBB ($ipSAE_{Actloop}$ = 0.64) [50] (**Figure 13**) which was not present in the template set for our structure prediction pipeline or AlphaFold2 training data. This is currently the only example of an experimentally determined active structure from this sub-family in the PDB. The key difference between 9m6pBBB and other PDB structures of this family (PKMYT1, WEE1, and WEE2) is the conformation of Glu-268 which is found at the APE10 position in the activation loop C-terminal segment of all three family members (E268, E477 and E394, equivalently). Glu[APE10] adopts an “APE10-in” conformation in our active models for all three kinases, as observed in 9m6pBBB (**Figure 13**). All other PDB structures of these kinases are found to adopt an “APE10-out” conformation, and we therefore classify them as inactive. It is likely that 9m6pBBB is representative of the active conformation of all three family members as predicted by our models

(**Figure 13**). We reiterate however that the activation loop prediction scores (**Equation 1**) for WEE1 and WEE2 are very low ($ipSAE_{Actloop}$ = 0.35), and the active conformation was only obtained using specific templates at very low MSA depths (three ortholog sequences each), similar to the LMTK family kinases.

DISCUSSION

With the advent of AlphaFold, much effort has gone into AF and similar programs to produce models of proteins and protein complexes in different states [20, 51-54]. Most of these do so by manipulating the multiple sequence alignment input to AF. For many proteins, standard AlphaFold may produce models in one state or another state but not both. Even if an ensemble of AF models can be generated with AlphaFold or its variants that fall into two or more conformational bins, it is important to be able to identify the functional characteristics of the different states for any downstream analysis. A key step in this process is establishing criteria for the desired state of the modeled proteins, especially for protein families of high biological and clinical relevance [55].

From a structural bioinformatics analysis of substrate bound typical protein kinase domains in the PDB, we have established criteria for identifying active experimental or predicted structures of protein kinases that are likely capable of binding ATP, metal ions, and substrates and catalyzing protein phosphorylation. The substrate-bound set is the largest ever compiled, consisting of 54 unique kinase-substrate pairs in 248 structures, including autophosphorylation structures that we identified in crystals [28]. Substrate binding has not typically been considered in previous efforts in identifying active structures of protein kinases.

Close inspection of the representative structures of each complex (**Table 1**) reveals substrate-binding surfaces formed by various structural elements. These structural elements are formed by residues of the catalytic loop (including the HRD motif) and residues at the N- and C-terminal ends of the activation loop (**Figure 2**). We show that the activation loop C-terminal motif (ActLoopCT) differs significantly between TYR and non-TYR kinases, and we therefore treat them somewhat differently (**Table 3**). We applied the full set of criteria to all experimental structures of kinases in the PDB and present these data on the updated Kincore website. In the PDB, as of January 21, 2026, there were active structures of 130 (29.8%) of the 437 human catalytic protein kinase domains.

Other kinase active/inactive classifiers have been developed in recent work using machine-learning [12, 56] , but none of these methods have explicit criteria for binding of substrates and they were trained on “active” kinase structures incapable of binding substrates. Consequently, these classifiers often fail to classify structures as inactive that have an activation loop conformation that would clearly interfere with the substrate binding modes observed in experimental structures of closely related kinases solved to-date: e.g., CAMK_AURKA (1mq4A), CMGC_CDK2 (1okvC), CAMK_DAPK1 (1jktA), CMGC_MAPK1 (4iz5A). Our classifier is instead comprised of joint structural criteria which are based on the structural bioinformatics of substrate-bound complexes.

From the substrate bound structures, supplemented with structures of other kinases that had very similar conformations of the activation loop, we compiled a template set that we used to model all 437 human protein kinase domains, in conjunction with multiple sequence alignments (typically 50 sequences) of diverse orthologs or close paralogs to each kinase. The same criteria enabled us to identify active structures among the models produced by AlphaFold2. We demonstrated that the models with the highest intramolecular ipSAE score of the activation loop also most closely resemble substrate-bound structures of kinases in the PDB, enabling us to choose the most suitable model for public distribution via the Kincore website. In all, we generated approximately 400,000 models to identify active model structures for all 437 catalytic human kinases.

The current set of active models for all 437 catalytic typical protein kinases in the human proteome are available on the KinCoRe database (https://dunbrack.fccc.edu/kincore/activemodels). The functionality of the website has been upgraded significantly, such that tables that result from searches display all structural parameters used to determine activity (on the "advanced" pages, e.g., https://dunbrack.fccc.edu/kincore/GENE/PRKACA?advanced=1. This is true for all experimental structures currently in the PDB as well as our models. We believe our models will be useful in understanding the structural basis of kinase substrate specificity, since they place kinase substrate-binding residues of the activation loop and the active site in suitable positions for catalysis.

Of course, the true active state may be better represented by ensembles of structures. These are available for many kinases by downloading the active structures in the PDB via the Kincore website search page. Furthermore, we believe our active kinase models for all 437 human catalytic kinases serve as reasonable starting structures for molecular dynamics simulations to explore the conformational space within the active energy basin. Notably, it is not uncommon to find that the starting structures for MD simulations (intended to sample the active basin) are not actually active. For example, the inactive structure of TYR_SRC (PDB: 1y57A) is often used as the basis of molecular dynamics simulations of the *active* protein [57-59]. But the structure of chicken SRC (3dqwA,C), which is only two residues different from human SRC, much more closely resembles the structure of substrate-bound structure of the closely related ABL1 kinase domain (2g2iB). There are large differences in the C-terminal end of the activation loop that make 1y57A inactive.

Future updates to our set of models will be possible as new kinase-substrate complexes are deposited in the PDB. This will be important especially for kinases that are currently lacking structural data on the substrate-bound state, e.g., dual-specificity kinases, non-plant kinases from the TKL family, kinases from the NEK family, and several distantly related kinases designated as OTHER. We plan to add ATP-bound complexes generated with Boltz2 [60] to Kincore, using templates generated in this work. Another challenge at the forefront of kinase structural modeling is generating models of kinases in complex with various substrates. For typical protein kinases, AlphaFold-Multimer is capable of modeling substrate-bound complexes when given a peptide substrate and Uniref90 as a sequence database, but

without the methods employed in this work (appropriate templates and carefully selected sequence alignments), it is not always capable of modeling the kinase activation loop in the active state. This will take additional study and implementation to develop a robust protocol that reliably makes models of kinase-substrate complexes from suitable choices of templates and multiple sequence alignments for AlphaFold-Multimer. This work is ongoing.

METHODS

Identifying substrate-bound structures in the PDB

Structures of kinases in complex with peptide/protein substrates (heterodimers and homodimers) were identified in the August 2024 PDB by checking each entry for intermolecular contacts between the HRD Asp delta oxygens and residues from chains in the same asymmetric unit or from neighboring asymmetric units, using the methods we described previously [28]. For crystal structures, the first step is to use crystal symmetry operators to build asymmetric units that are in contact with any protein atoms in the deposited structure. These can be in the same unit cell or neighboring unit cells. A generous heavy-atom distance cutoff of 4.5 Å was used to identify contacts of side-chain hydroxyls of potential substrates in one protein chain and the HRD Asp of any kinase chain. This search yielded a set of candidate transphosphorylation and autophosphorylation structures. As before, we compared these potential phosphorylation complexes with known peptide-substrate-bound protein kinase structures. This enables us to eliminate domain-swapped activation loop dimers, because the swapped Thr side chain in these structures approaches the HRD-Asp from below, rather than from above [28]. It is in the same position both in sequence and structure as the intramolecular hydrogen bond of APE8.

Some structures were not identified automatically by the structural search and needed to be added manually to the list of substrate-bound complexes. For example, TYR_ABL1 has two bisubstrate-complexes which have been long recognized as inactive (2g2f: DFG-out and 2g1t: “Src-like inactive”). However, the substrate-mimic structure (P0 Tyr→Phe mutant) of TYR_ABL1 bound to ADP (2g2i) has an activation loop conformation that resembles typical substrate-bound TYR structures, though the substrate-mimic Phe in each of these structures has a distance greater than 4.5 Å from the HRD Asp. 2g2i appears to be a reasonable model of the substrate-bound state of ABL1, except for the HRD motif which is HRD-out. The structure of OTHER_CDC7 bound to bisubstrate (a peptide covalently linked to an ATP analog) has historically served as a putative model of the substrate-bound state for this kinase (6ya7) and is corroborated by the structures of other Ser/Thr-specific substrate-bound kinases which display a similar “upside-down boat” conformation of the ActLoopCT. This complex was not automatically identified due to a non-standard residue code of the substrate mimic residue.

Some complexes appear many times in the PDB (e.g., CAMK_PIM1 with pimtide, OTHER_PINK1 autophosphorylation complex) which obscures the true number of unique, experimentally observed complexes. Based on the identity of enzyme kinase and substrate, we counted a total of 54 unique

complexes from this dataset. Representative structures of each complex were chosen for visualization in **Figure 2** – for this we selected structures bound to ATP analogs and with complete coordinates for the activation loop, when available.

Some of the structures in **Table 1** contain kinases bound to folded proteins, and not peptides. These are either autophosphorylation complexes or bound protein substrates or inhibitor proteins. The autophosphorylation complexes involve sites in the activation loop (STE_PAK1, 4zy4; TYR_IGF1R, 3lvp; TKL_IRAK4, 4u97 and 8wtf; TYR_FGFR3, 8udu), the kinase insert loop (TYR_FGFR1, 3gqi; TYR_FGFR3, 4k33), and N or C terminal tails (CAMK_CAMK2D, 2wel; CAMK_CAMKII, 3kk8; TYR_CSF1R, 3lcd; TYR_KIT, 1pkg; TYR_EPHA2, 4pdo; TYR_JAK2, 6vnb; TYR_FGFR2, 3cly and 2pvf). The mitochondrial Ser/Thr kinase OTHER_PINK1 autophosphorylates on the β3-αC loop [61] as shown in 8uyh. For the activation loop autophosphorylation structures of TKL_IRAK4, we note the presence of both a symmetric (8wtf) and asymmetric (4u97) dimer in the PDB – this is a unique example where the activation loop of the substrate-bound enzyme adopts two different conformations, and it is likely that the asymmetric dimer 4u97, which has a more-compact activation loop structure of the enzyme monomer (resembling the fully phosphorylated structure: 8w3w), is the more biologically relevant of the two dimers. Three other complexes are with larger proteins which are either direct substrates or substrate-mimicking inhibitors or both (AGC_PRKACA:RYR2, 6mm7; AGC_PRKACA:KAP3, 3tnp; TKL_BAK1:HPAB2, 3tl8). The last of these is a plant kinase/pathogen-inhibitor complex [62]. These examples provide insights of how kinases phosphorylate amino acids in the context of folded protein domains, as opposed to intrinsically disordered regions (IDRs) [28].

Ortholog sequence sets

We first searched UniProt for Pfams PF00069 and PF07714 to collect a set of 1.68 million sequences in UniRef100 with typical protein kinase domains. For each of 437 catalytic kinase domain sequences from our earlier alignment of all human kinase domains [3], we used PSI-BLAST [63] to get a list of the top 25,000 closest kinases to each human kinase domain. The queries used were 8 residues longer on each end of the kinase domain than our published alignment. The hit regions in the PSI-BLAST output were then filtered for sequences more than 50% identical to the query, coverage greater than 90% of the query length, and gap percentage in the alignment of less than 10%. We then applied CD-HIT [45] to create lists of orthologs (or close paralogs) with no more than 90% sequence identity to each other. These sequences were used as query databases to generate MSAs for our AlphaFold2 calculations.

AlphaFold2/Colabfold

To predict the structures of human catalytic kinases we used the Colabfold implementation (version 1.5.0) of DeepMind's advanced machine learning model, AlphaFold2 [18, 46]. The code was downloaded from the official Colabfold GitHub repository (https://github.com/sokrypton/ColabFold) and

used to generate protein structures on a Linux-based high performance computing cluster equipped with NVIDIA L40S GPUs.

Data Input and Preparation: Three sets of sequence databases were used to create multiple sequence alignments: the default UniRef30 database (via the MMseqs2 server), an additional kinase family-focused sequence database (all 493 human kinases in the human proteome, separated into each family), and a kinase orthologs-focused sequence database (described above). Templates for the calculations were obtained from the August 2024 PDB set by performing an all-vs-all superposition of all kinase structures determined as active against each of the 248 substrate-bound structures, allowing us to identify active structures which are also within 1.5 Å RMSD of the substrate-bound state (based on the first 9 and last 15 residues of the activation loop), prioritizing templates which are themselves bound to substrate or an ATP analog. From this we acquired 101 template structures, limited to one template per-kinase to avoid over-representing certain kinase sequences in the template set. For each of the 437 human catalytic kinases, we ran AlphaFold2 using the 101 template structures individually (using a single chain for each template).

Model Configuration and Implementation: Calculations with AlphaFold2 were conducted using Colabfold[46]. The following command was used:

```
colabfold_batch --model-type alphafold2_multimer_v3 --num-seeds 1 --num-models 5 --
templates --custom-template-path template_folder a3m_msa_input_file output_folder/
```

For model generation without templates using the default MSA from MMseqs2, the following command was used:

```
colabfold_batch --model-type alphafold2_multimer_v3 --num-seeds 101 --num-models 5
--templates fasta_sequence_file output_folder/
```

The multiple sequence alignment was prepared using the hh-suite package [64] and subsequently fed into the model for structure prediction. When using templates, we ran one seed using one template at a time. When using no templates, we ran Colabfold using default settings which uses MMseqs2 and the UniRef30 database [47] to generate the input sequence alignments, and 101 seeds to match the sampling provided by the template procedure.

We ran Colabfold with specific sequence data sets by providing our own custom sequence databases at runtime: Ortholog and Family for the MSA building step and individual templates to predict protein structures. When predicting the structure of a given target, we did not use “self-templates” i.e., templates of the target kinase were avoided in the template set (resulting in 100 templates deployed per-kinase target, down from 101).

Benchmarking

The predicted structures were validated by comparing them to the benchmark PDB structures of kinases. The validation process relied on our ipSAE score [22] defined intramolecularly between the activation loop and kinase domain (**Equation 1**) and Root Mean Square Deviation (RMSD), measuring the distance between atoms in the predicted and known structures. Model structures were aligned to benchmark structures using the SVDSuperimposer() method in BioPython [65]. The alignment was performed on the C-terminal domain of each structure (40 residues C-terminal to the APE motif). RMSD was measured for the activation loop backbone atoms (N, CA, C, O) after superposition of the C-terminal domains.

Two benchmarks were constructed using the August 2024 PDB set of human kinases (plus non-human orthologs). The first benchmark contained substrate-bound structures from Table 1 with complete coordinates for the activation loop in the PDB structure (29 kinases). The second benchmark consisted of 117 structures with complete (or mostly complete) activation loops that passed our active criteria from the PDB.

While constructing the second benchmark, most kinases had multiple structures to choose from. The structure with the lowest RMSD to one of the substrate-bound structures was chosen. Structures bound to ATP or ATP analogs were prioritized when available. This resulted in a set of mostly human kinases with complete activation loops and some non-human orthologs. Only two structures from this set were missing coordinates in the activation loop: the benchmark structure for OTHER_AAK1 is missing two residues in the middle the loop (Q206 and A207) between DFG12 and APE22. The benchmark structure for TKL_LIMK1 is missing a ~16 residue-long insertion in the middle of the activation loop between DFG8 and APE18, and this part of the loop may not have a well-defined structure even when bound to substrate. In both cases, the missing residues are outside the interaction region for substrates. We note that activation loop contacts between non-TYR kinases and bound substrates drop off almost entirely at DFG7 and APE15 (Figure 4).

A third benchmark was constructed post hoc using the January 2026 PDB, which provided active structures for two human kinases not in the August 2024 set: OTHER_PKMYT1 (9m6pBBB) and CMGC_MAPK14 (9cj2A): both were found to be within 1.3 Å of the active AlphaFold2 model.

Update to Kincore Database and Website and Kincore_Standalone3

We have updated the Kincore database and website (https://dunbrack.fccc.edu/kincore) to include the additional active criteria defined in this study (Salt bridge, HRD, ActLoopNT, ActLoopCT) for all structures in the PDB. Each structure is marked “Active” or “Inactive” or “None” (missing atoms) based on these criteria. Pseudokinases are marked “Pseudo.” We have also added the highest scoring AlphaFold2 active model for each of the human catalytic kinases to Kincore. These are labeled with the

prefix “AF-” and the suffix “-K3” or “-K4” attached to the Uniprot Accession ID (e.g., the CAMK_AURKA model is AF-O14965-K3). “K4” is used for the second kinase domain of kinases that contain two kinase domains. Otherwise “K3” is used. “K1 and K2” were used for the models we produced earlier but are now marked “Active” or “Inactive” by the updated criteria presented in this paper. The website now allows the user to search for “Active”, “Inactive”, and “Pseudo” kinase structures on the main search page (https://dunbrack.fccc.edu/kincore/formSearch). We have also added a ligand-type called “ATPlike” which includes structures bound with ATP and its analogs (ACP, ANP, ADP, ACS). The updated standalone program for assessing the conformational state of protein kinases, Kincore-standalone3, is available at https://github.com/DunbrackLab/Kincore-standalone3/ .

Pseudocode that describes the algorithm/pipeline used in Kincore-standalone3 is presented in the Supplementary Methods.

Data Availability and Reproducibility

To ensure the reproducibility of our study, all data, including input sequences, ortholog sequence sets, and predicted structures, are accessible at http://dunbrack.fccc.edu/kincore/activemodels. PDB reference codes and RCSB links for all structures referenced in this paper are as follows:

pdb_00001fin, pdb_00001gag, pdb_00001jkt, pdb_00001k3a, pdb_00001mq4, pdb_00001o6k, pdb_00001okv, pdb_00001ol5, pdb_00001pkg, pdb_00001qmz, pdb_00001vyw, pdb_00001x8b, pdb_00001y57, pdb_00001yi6, pdb_00002bkz, pdb_00002bpm, pdb_00002bzk, pdb_00002c4g, pdb_00002erk, pdb_00002g2i, pdb_00002j50, pdb_00002phk, pdb_00002pvf, pdb_00002q0n, pdb_00002wel, pdb_00002wih, pdb_00002wo6, pdb_00002wpa, pdb_00002xuu, pdb_00003bu5, pdb_00003cd3, pdb_00003cly, pdb_00003dgk, pdb_00003dqw, pdb_00003e5a, pdb_00003eoc, pdb_00003erk, pdb_00003fxx, pdb_00003gqi, pdb_00003ha6, pdb_00003kk8, pdb_00003lcd, pdb_00003lvp, pdb_00003mdy, pdb_00003qhr, pdb_00003tl8, pdb_00003tnp, pdb_00003x2u, pdb_00004dee, pdb_00004ekk, pdb_00004iz5, pdb_00004jdh, pdb_00004k33, pdb_00004ouc, pdb_00004pdo, pdb_00004r3p, pdb_00004rio, pdb_00004u97, pdb_00004x7q, pdb_00004xj0, pdb_00004zy4, pdb_00005c27, pdb_00005czh, pdb_00005czi, pdb_00005dt3, pdb_00005eg3, pdb_00005fg8, pdb_00005fri, pdb_00005g1x, pdb_00005li1, pdb_00005lih, pdb_00005lxm, pdb_00005oro, pdb_00005os0, pdb_00005v60, pdb_00005vdk, pdb_00005w84, pdb_00006cqd, pdb_00006e2m, pdb_00006i2u, pdb_00006mm5, pdb_00006mm7, pdb_00006r49, pdb_00006r4d, pdb_00006ru7, pdb_00006vnb, pdb_00006vpg, pdb_00006w3b, pdb_00006w3c, pdb_00006wlx, pdb_00006wly, pdb_00006ya7, pdb_00007e12, pdb_00007s46, pdb_00007s47, pdb_00007ujp, pdb_00007w5o, pdb_00007xzr, pdb_00007ztl, pdb_00008d7n, pdb_00008pqh, pdb_00008qgy, pdb_00008udu, pdb_00008uyh, pdb_00008wtf, pdb_00008yhw.

ACKNOWLEDGMENTS

This work was funded by NIH Grant R35 GM122517 (to RLD) and P30 CA006927 (to Fox Chase Cancer Center). Calculations were performed on the Fox Chase High Performance Computer Cluster.

REFERENCES

1 Cohen, P., Cross, D. and Jänne, P.A. (2021) Kinase drug discovery 20 years after imatinib: progress and future directions. *Nature reviews drug discovery* **20**, 551-569

2 Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Ashford, V.A., Xuong, N.-H., Taylor, S.S. et al. (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407-414

3 Modi, V. and Dunbrack, R.L., Jr. (2019) A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. *Scientific reports* **9**, 19790 10.1038/s41598-019-56499-4

4 Modi, V. and Dunbrack, R. (2022) Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Res.* **50**, D654-D664

5 Jacobs, M.D., Caron, P.R. and Hare, B.J. (2008) Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: structure of LCK/imatinib complex. *Proteins* **70**, 1451-1460

6 Hari, S.B., Merritt, E.A. and Maly, D.J. (2013) Sequence determinants of a specific inactive protein kinase conformation. *Chem. Biol.* **20**, 806-815 10.1016/j.chembiol.2013.05.005

7 Möbitz, H. (2015) The ABC of protein kinase conformations. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1854**, 1555-1566

8 McSkimming, D.I., Rasheed, K. and Kannan, N. (2017) Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics* **18**, 86

9 Ung, P.M.-U., Rahman, R. and Schlessinger, A. (2018) Redefining the protein kinase conformational space with machine learning. *Cell chemical biology* **25**, 916-924. e912

10 Modi, V. and Dunbrack, R.L. (2019) Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences* **116**, 6818-6827 10.1073/pnas.1814279116

11 Kanev, G.K., de Graaf, C., Westerman, B.A., de Esch, I.J. and Kooistra, A.J. (2021) KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.* **49**, D562-D569

12 Reveguk, I. and Simonson, T. (2024) Classifying protein kinase conformations with machine learning. *Protein Sci.* **33**, e4918

13 Johnson, L.N., Noble, M.E. and Owen, D.J. (1996) Active and inactive protein kinases: structural basis for regulation. *Cell* **85**, 149-158

14 Levinson, N.M., Kuchment, O., Shen, K., Young, M.A., Koldobskiy, M., Karplus, M. et al. (2006) A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol.* **4**, e144

15 Schindler, T., Bornmann, W., Pellicena, P., Miller, W.T., Clarkson, B. and Kuriyan, J. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science* **289**, 1938-1942

16 Yang, J., Wu, J., Steichen, J.M., Kornev, A.P., Deal, M.S., Li, S. et al. (2012) A conserved Glu–Arg salt bridge connects coevolved motifs that define the eukaryotic protein kinase fold. *J. Mol. Biol.* **415**, 666-679

17 Lowe, E.D., Noble, M.E., Skamnaki, V.T., Oikonomakos, N.G., Owen, D.J. and Johnson, L.N. (1997) The crystal structure of a phosphorylase kinase peptide substrate complex: kinase substrate recognition. *EMBO J.* **16**, 6646-6658 10.1093/emboj/16.22.6646

18 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589

19 Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G. et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439-D444

20 Del Alamo, D., Sala, D., Mchaourab, H.S. and Meiler, J. (2022) Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751

21 Heo, L. and Feig, M. (2022) Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins: Structure, Function, and Bioinformatics* **90**, 1873-1885

22 Dunbrack Jr, R.L. (2025) Rēs ipSAE loquuntur: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv* **https://doi.org/10.1101/2025.02.10.637595**

23 Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J.E., Mudge, J.M. et al. (2023) GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942-D949

24 Suijkerbuijk, S.J., van Dam, T.J., Karagöz, G.E., von Castelmur, E., Hubner, N.C., Duarte, A.M. et al. (2012) The vertebrate mitotic checkpoint protein BUBR1 is an unusual pseudokinase. *Dev. Cell* **22**, 1321-1329

25 Katso, R., Russell, R. and Ganesan, T. (1999) Functional analysis of H-Ryk, an atypical member of the receptor tyrosine kinase family. *Mol. Cell. Biol.* **19**, 6427-6440

26 Seal, R.L., Braschi, B., Gray, K., Jones, T.E., Tweedie, S., Haim-Vilmovsky, L. et al. (2023) Genenames. org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003-D1009

27 Lin, K., Lin, J., Wu, W.-I., Ballard, J., Lee, B.B., Gloor, S.L. et al. (2012) An ATP-site on-off switch that restricts phosphatase accessibility of Akt. *Science signaling* **5**, ra37-ra37

28 Xu, Q., Malecka, K.L., Fink, L., Jordan, E.J., Duffy, E., Kolander, S. et al. (2015) Identifying three-dimensional structures of autophosphorylation complexes in crystals of protein kinases. *Sci Signal* **8**, rs13 https://doi.org/10.1126/scisignal.aaa6711

29 Hayward, S. (2001) Peptide-plane flipping in proteins. *Protein Sci.* **10**, 2219-2227

30 Kornev, A.P. and Taylor, S.S. (2010) Defining the conserved internal architecture of a protein kinase. *Biochim Biophys Acta* **1804**, 440-444 10.1016/j.bbapap.2009.10.017

31 Touw, W.G., Joosten, R.P. and Vriend, G. (2015) Detection of trans–cis flips and peptide-plane flips in protein structures. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **71**, 1604-1614

32 Meyder, A., Nittinger, E., Lange, G., Klein, R. and Rarey, M. (2017) Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *Journal of chemical information and modeling* **57**, 2437-2447

33 Shapovalov, M., Vucetic, S. and Dunbrack, R.L., Jr. (2019) A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLOS Comput. Biol.* **15**, e1006844

34 Shapovalov, M.V. and Dunbrack, R.L., Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844-858

35 Dunbrack, R.L., Jr. (1993) Conformational analysis of protein sidechains: Derivation of empirical energy parameters for proline and development of a backbone-dependent rotamer library. PhD Dissertation, Harvard University, Cambridge, MA

36 Schmidt, J.M., Brueschweiler, R., Ernst, R.R., Dunbrack Jr, R.L., Joseph, D. and Karplus, M. (1993) Molecular dynamics simulation of the proline conformational equilibrium and dynamics in antamanide using the CHARMM force field. *Journal of the American Chemical Society* **115**, 8747-8756

37 Ho, B.K. and Brasseur, R. (2005) The Ramachandran plots of glycine and pre-proline. *BMC Struct. Biol.* **5**, 14

38 Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M.I. and Dunbrack, R.L., Jr. (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLOS Comput. Biol.* **6**, e1000763

39 Bayliss, R., Sardon, T., Vernos, I. and Conti, E. (2003) Structural basis of Aurora-A activation by TPX2 at the mitotic spindle. *Mol. Cell* **12**, 851-862

40 Richards, M.W., Burgess, S.G., Poon, E., Carstensen, A., Eilers, M., Chesler, L. et al. (2016) Structural basis of N-Myc binding by Aurora-A and its destabilization by kinase inhibitors. *Proceedings of the National Academy of Sciences* **113**, 13726-13731

41 Derewenda, Z.S., Lee, L. and Derewenda, U. (1995) The occurence of C–H⋯ O hydrogen bonds in proteins. *J. Mol. Biol.* **252**, 248-262

42 Modi, V. and Dunbrack, R.L. (2019) A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. *Scientific Reports* **9**, 1-16 10.1038/s41598-019-56499-4

43 Gizzio, J., Thakur, A., Haldane, A., Post, C.B. and Levy, R.M. (2024) Evolutionary sequence and structural basis for the distinct conformational landscapes of Tyr and Ser/Thr kinases. *Nat Commun* **15**, 6545 10.1038/s41467-024-50812-0

44 Oruganty, K., Talathi, N.S., Wood, Z.A. and Kannan, N. (2013) Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. *Proceedings of the National Academy of Sciences* **110**, 924-929

45 Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 10.1093/bioinformatics/bts565

46 Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. and Steinegger, M. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679-682

47 Steinegger, M. and Soding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 10.1038/nbt.3988

48 Luz, S., Cihil, K.M., Brautigan, D.L., Amaral, M.D., Farinha, C.M. and Swiatecka-Urban, A. (2014) LMTK2-mediated phosphorylation regulates CFTR endocytosis in human airway epithelial cells. *J Biol Chem* **289**, 15080-15093 10.1074/jbc.M114.563742

49 Squire, C.J., Dickson, J.M., Ivanovic, I. and Baker, E.N. (2005) Structure and Inhibition of the Human Cell Cycle Checkpoint Kinase, Wee1A Kinase: An AtypicalTyrosine Kinase with a Key Role in CDK1 Regulation. *Structure* **13**, 541-550

50 Wang, Y., Wang, X., Liu, T., Wang, C., Meng, Q., Meng, F. et al. (2025) Discovery of a bifunctional PKMYT1-targeting PROTAC empowered by AI-generation. *Nat Commun* **16**, 10759 10.1038/s41467-025-65796-8

51 Kalakoti, Y. and Wallner, B. (2024) AFsample2: Predicting multiple conformations and ensembles with AlphaFold2. *BioRxiv*, 2024.2005. 2028.596195

52 Wayment-Steele, H.K., Ojoawo, A., Otten, R., Apitz, J.M., Pitsawong, W., Hömberger, M. et al. (2024) Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832-839

53 Monteiro da Silva, G., Cui, J.Y., Dalgarno, D.C., Lisi, G.P. and Rubenstein, B.M. (2024) High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nature communications* **15**, 2464

54 Lee, M., Schafer, J.W., Prabakaran, J., Chakravarty, D., Clore, M.F. and Porter, L.L. (2025) Large-scale predictions of alternative protein conformations by AlphaFold2-based sequence association. *Nature communications* **16**, 5622

55 Parker, M.I., Meyer, J.E., Golemis, E.A. and Dunbrack, R.L., Jr. (2022) Delineating The RAS Conformational Landscape. *Cancer Res.* **In Press**

56 El Moumni, M. (2025) Resolving Discrepancies in Kinase Activity Labels Using Machine Learning. *Journal* 202506.200609 DOI:10.20944/preprints202506.0609.v1

57 Foda, Z.H., Shan, Y., Kim, E.T., Shaw, D.E. and Seeliger, M.A. (2015) A dynamically coupled allosteric network underlies binding cooperativity in Src kinase. *Nature communications* **6**, 5939

58 Fajer, M., Meng, Y. and Roux, B. (2017) The activation of c-Src tyrosine kinase: conformational transition pathway and free energy landscape. *The Journal of Physical Chemistry B* **121**, 3352-3363

59 Joshi, M.K., Burton, R.A., Wu, H., Lipchik, A.M., Craddock, B.P., Mo, H. et al. (2020) Substrate binding to Src: A new perspective on tyrosine kinase substrate recognition from NMR and molecular dynamics. *Protein Sci.* **29**, 350-359

60 Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V.R. et al. (2025) Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 10.1101/2025.1106.1114.659707

61 Rasool, S., Veyron, S., Soya, N., Eldeeb, M.A., Lukacs, G.L., Fon, E.A. et al. (2022) Mechanism of PINK1 activation by autophosphorylation and insights into assembly on the TOM complex. *Mol Cell* **82**, 44-59.e46 10.1016/j.molcel.2021.11.012

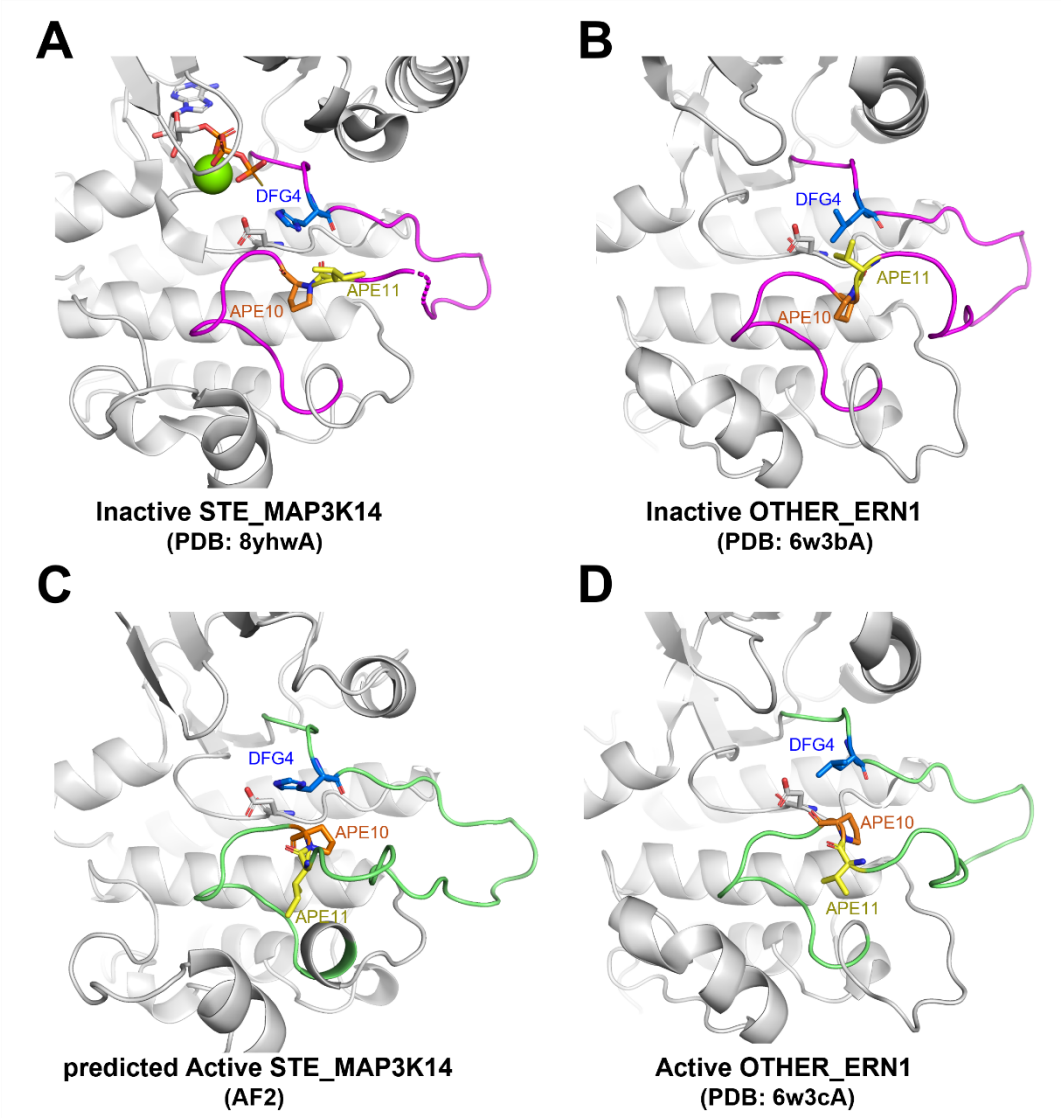
62 Cheng, W., Munkvold, K.R., Gao, H., Mathieu, J., Schwizer, S., Wang, S. et al. (2011) Structural analysis of Pseudomonas syringae AvrPtoB bound to host BAK1 reveals two similar kinase-interacting domains in a type III Effector. *Cell Host Microbe* **10**, 616-626 10.1016/j.chom.2011.10.013

63 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of database programs. *Nucleic Acids Res.* **25**, 3389-3402

64 Steinegger, M., Meier, M., Mirdita, M., Vohringer, H., Haunsberger, S.J. and Soding, J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 10.1186/s12859-019-3019-7

65 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A. et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423

Supplementary Material



Supplementary Figure 1. Justification for designating STE_MAP3K14 structures as inactive.

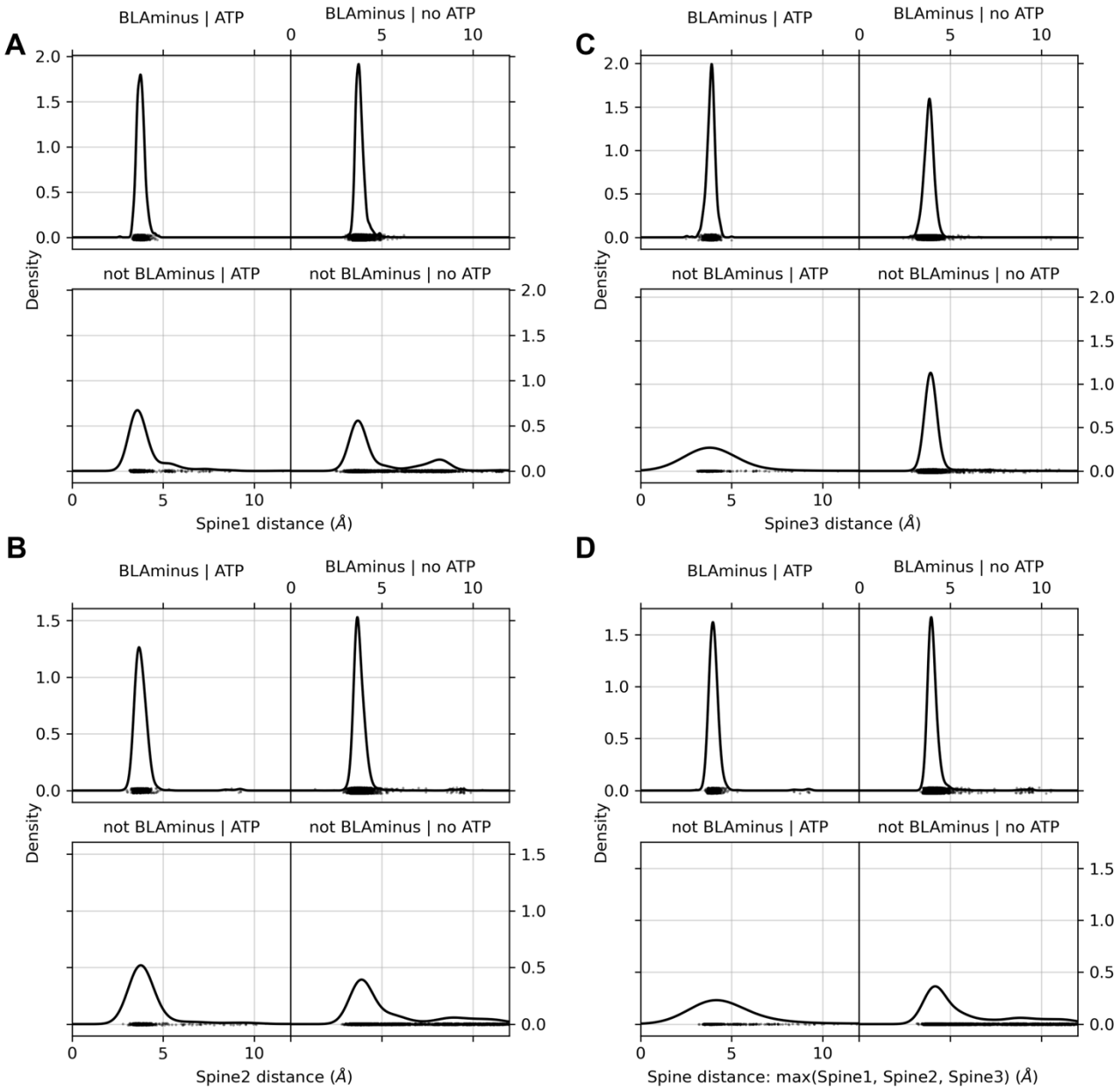
Activation loops of inactive structures are colored magenta and those of active structures are colored green. The HRD Asp residue and ATP analogs (where applicable) are shown with grey sticks.

A. Inactive STE_MAP3K14 structures have a contracted APE11-DFG4 distance that results from a flipped conformation of the Ile[APE11]-Pro[APE10] (I556 and P557) residues. We label all experimental MAP3K14 structures to-date as inactive due to this feature.

B. Inactive/unphosphorylated OTHER_ERN1 structures also have a contracted APE11-DFG4 distance which results from a flipped conformation of the Val[APE11]-Pro[APE10] (V731 and P732) residues, similar to MAP3K14.

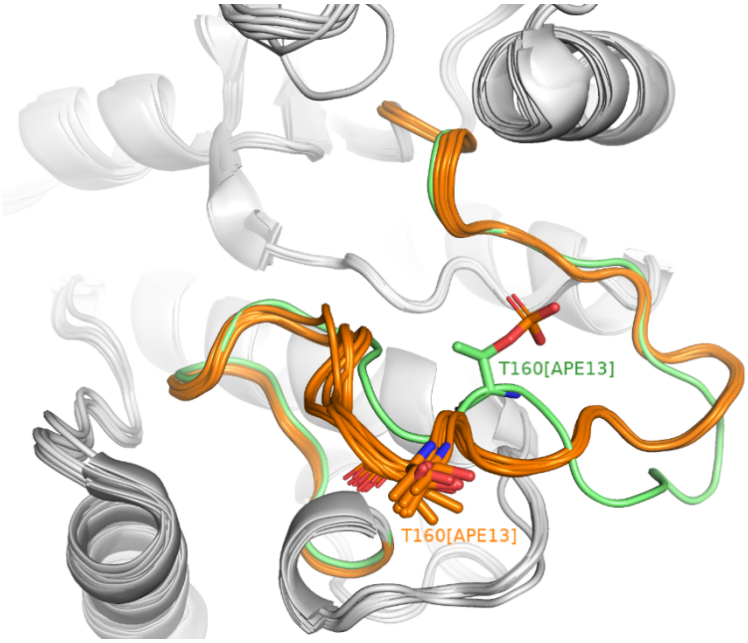
C. AlphaFold2 model of STE_MAP3K14 generated using Colabfold with no templates (see *Methods*) and $ipSAE_{Actloop}$ = 0.65 (see Equation 1 in main text), which is the highest-scoring model of STE_MAP3K14 generated in this study that can be classified as active. This model is also displayed in Figure 10 of the main text alongside other STE family members.

D. An active/phosphorylated experimental structure of OTHER_ERN1 (PDB: 6w3cA) displays the typical upside-down round-bottom boat structure in its ActLoopCT that is common for non-TYR kinases despite having Pro at APE10, suggesting that MAP3K14 should also be able to adopt a similar active conformation consistent with the predicted Active MAP3K14 structure in **C**.



Supplementary Figure 2. Distribution of regulatory spine distances.

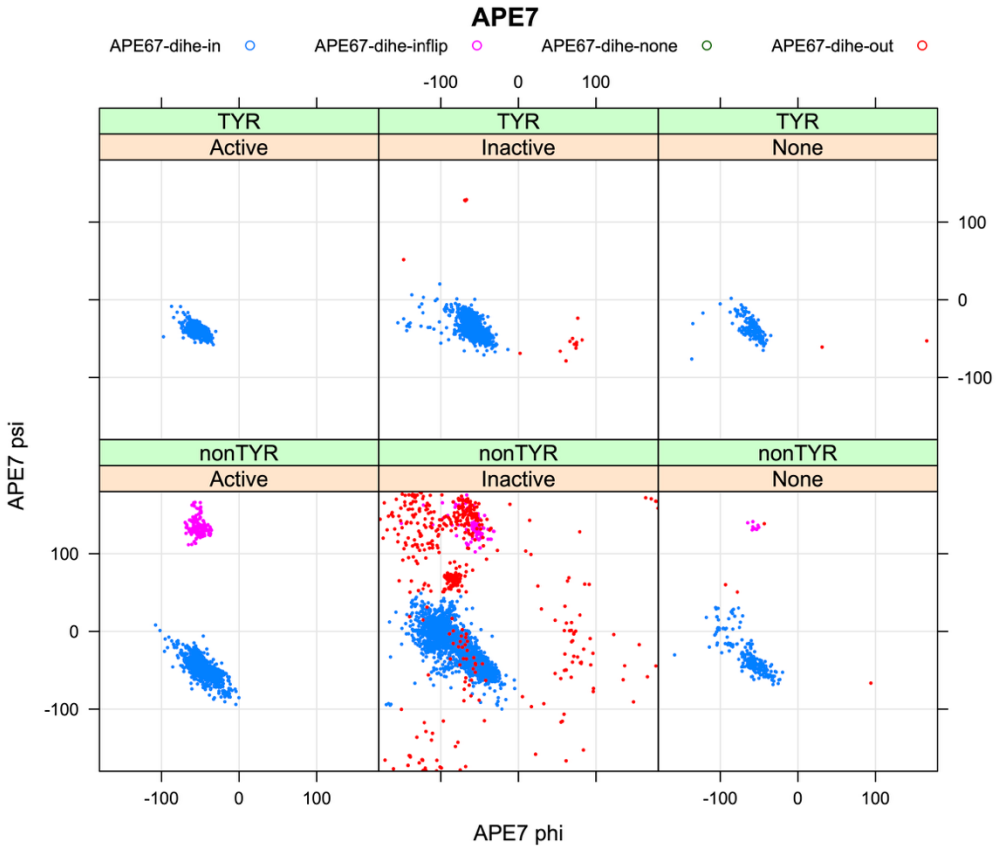
The Spine distances are defined as the closest distance among all side-chain atom pairs between the two residues. In each plot, there are 665 *BLAminus*/ATP points, 5043 *BLAminus*/no_ATP points, 454 not_*BLAminus*/ATP points, and 4316 not_*BLAminus*/no_ATP points.

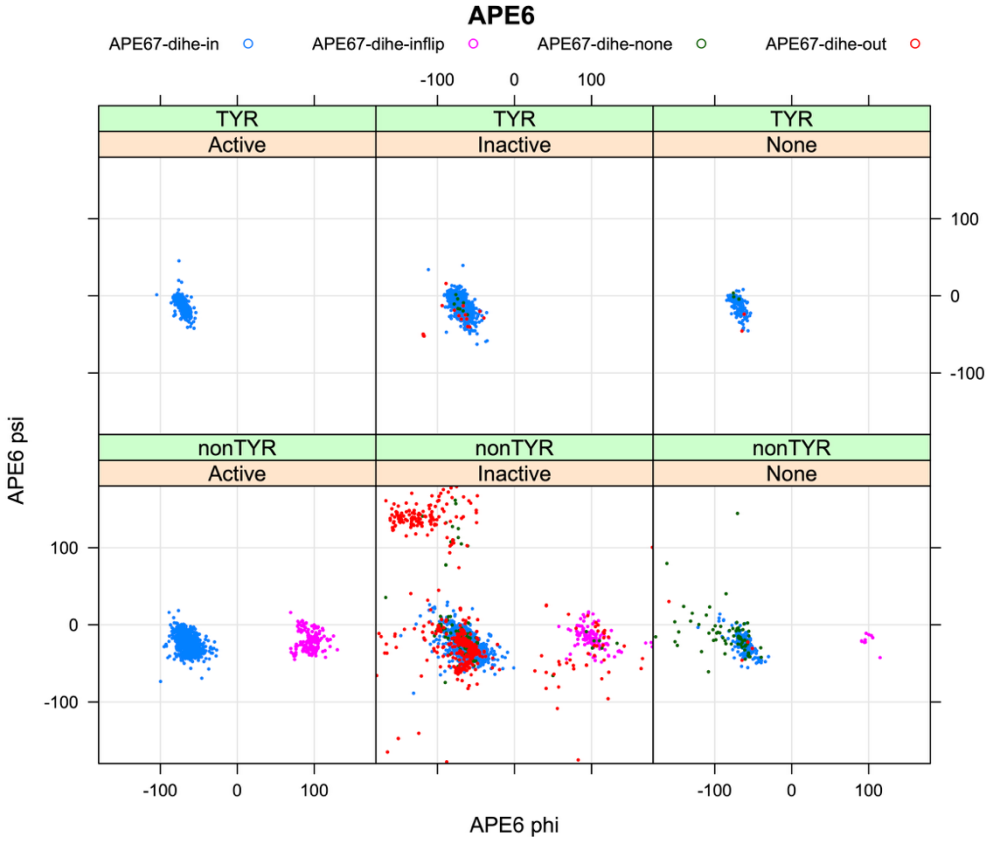


Supplementary Figure 3. Outlier structures of CMGC_CDK2 that pass our active criteria.

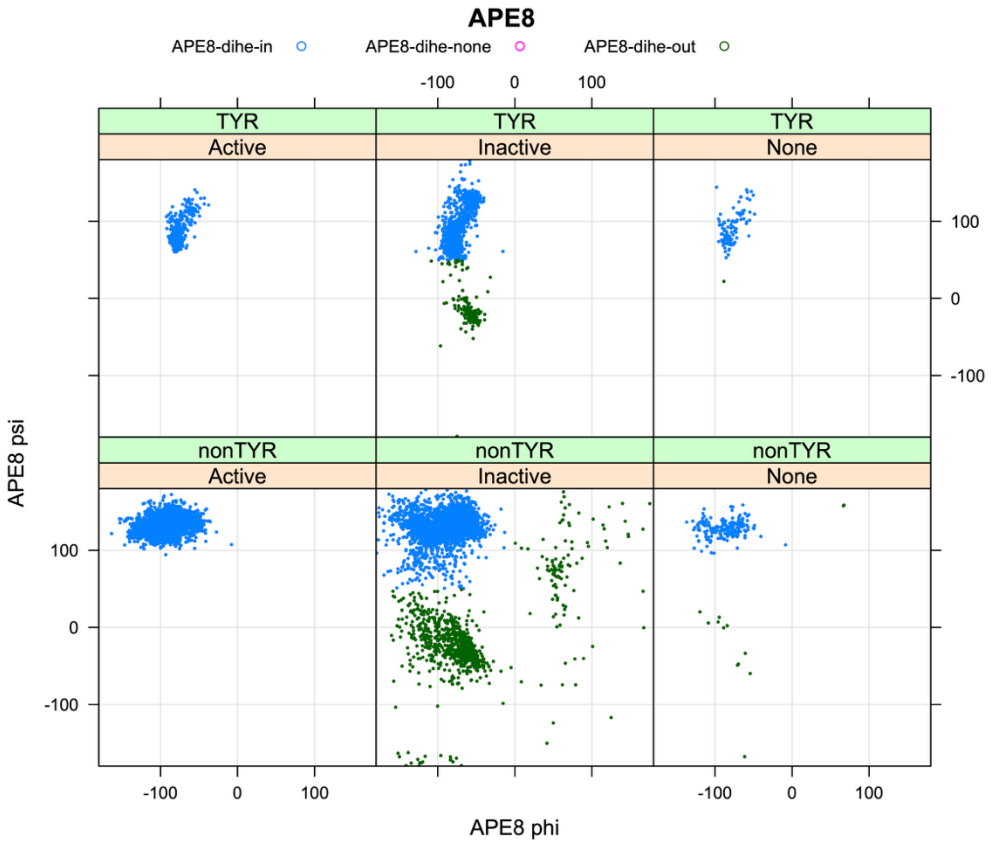
Among CMGC_CDK2 structures which are active by our criteria, 6% have activation loops that deviate from the substrate-bound structure by more than 2 Å RMSD. These outlier activation loops are colored orange (2c4gC, 2bpmC, 2wihC, 1vywC, 2wpaC, 2bkzC, 2bkzA, 3eocA), and the substrate-bound activation loop is colored green (1qmzA). In the outlier structures, T160[APE13] is unphosphorylated and flipped “out”.

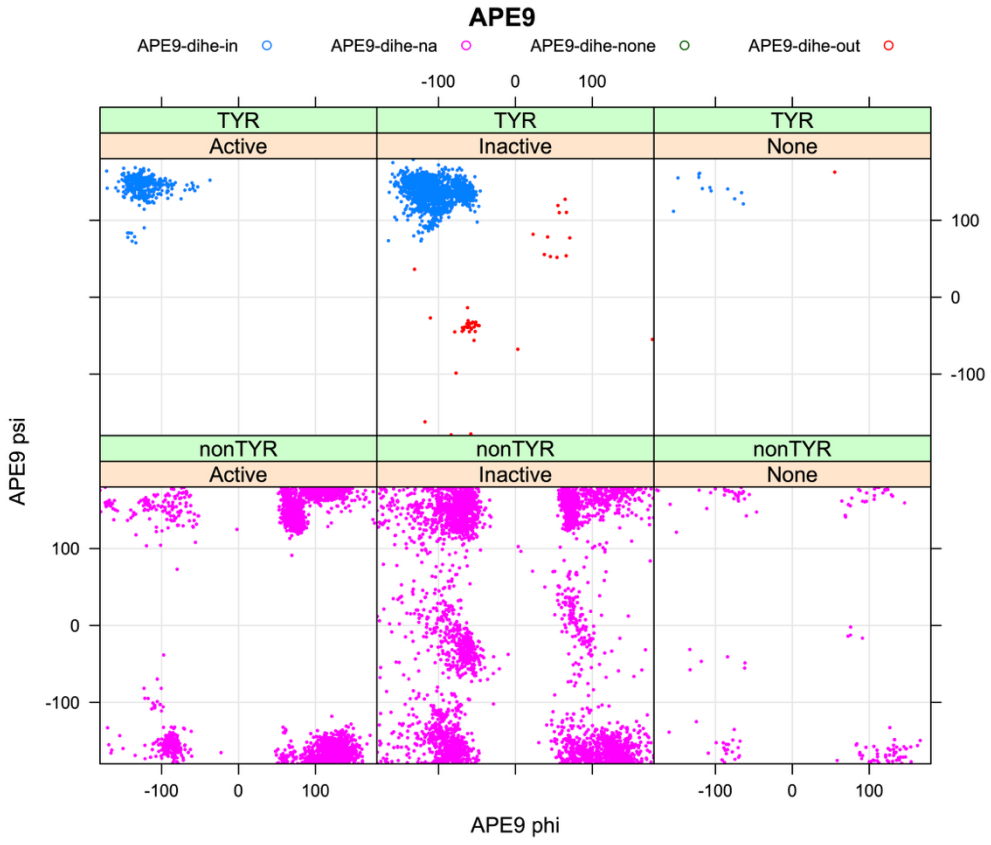
(Supplementary Figure 4)



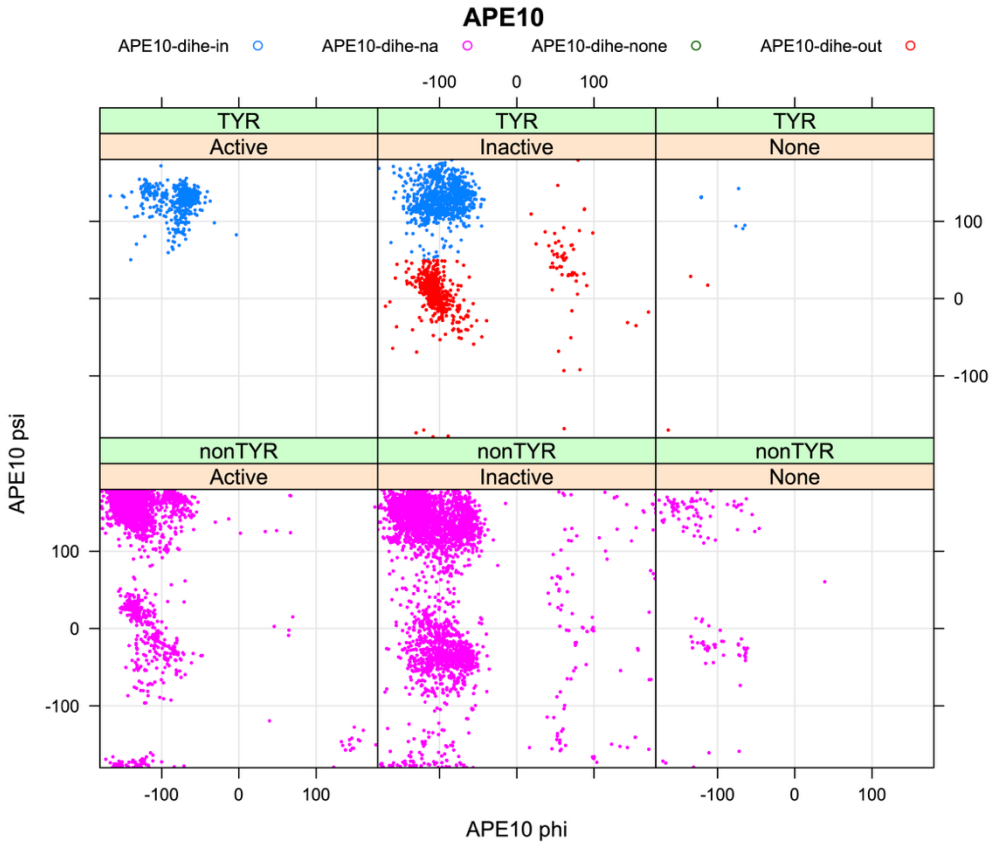


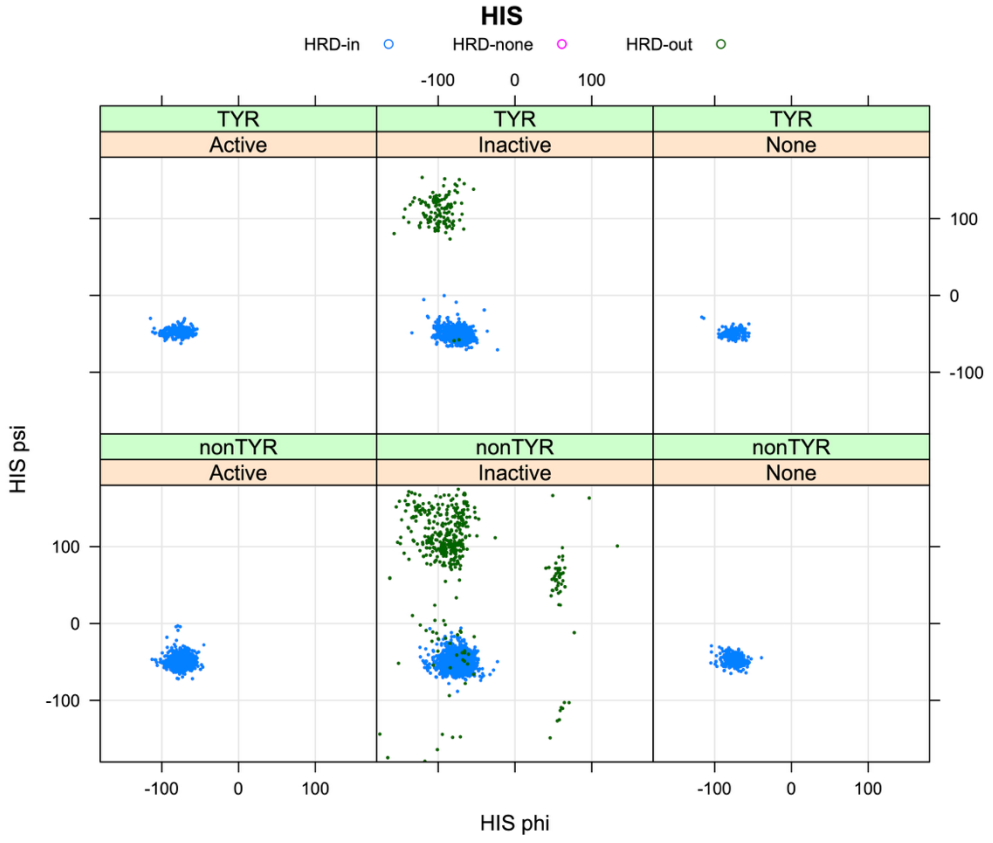
(Supplementary Figure 4, continued)

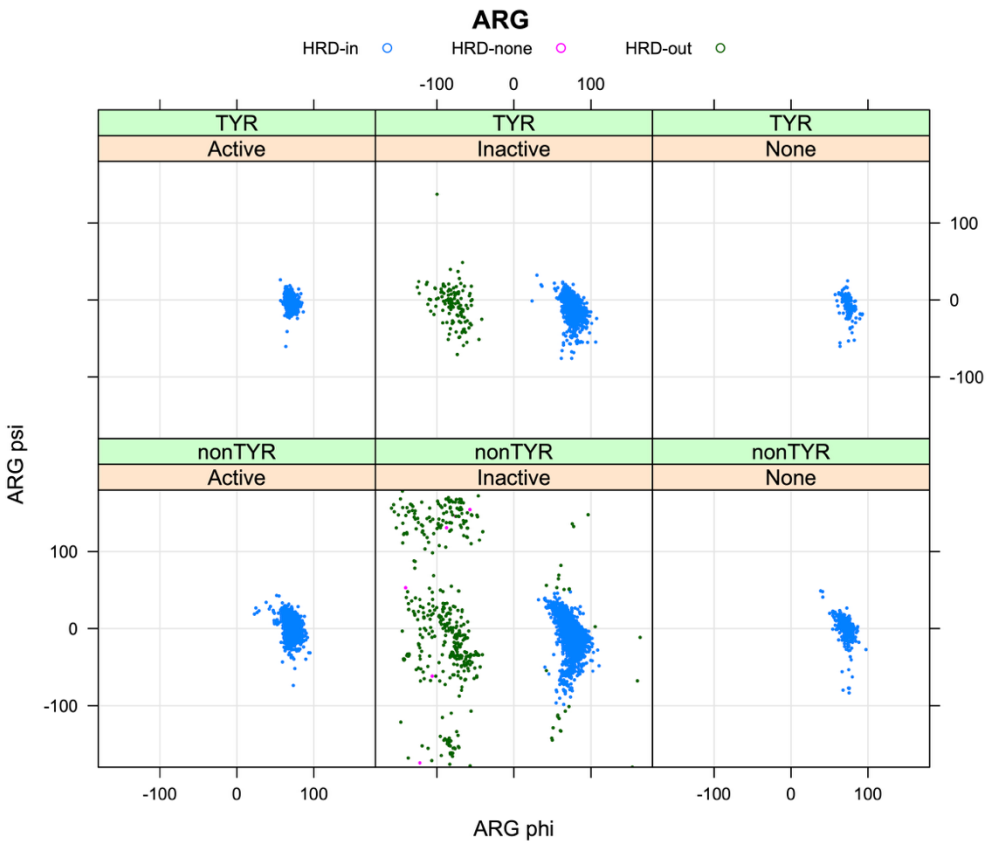




(Supplementary Figure 4, continued)

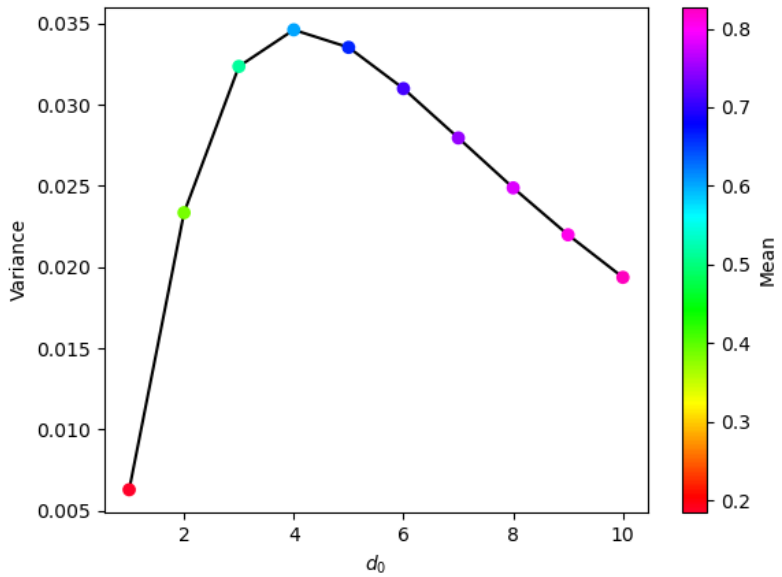






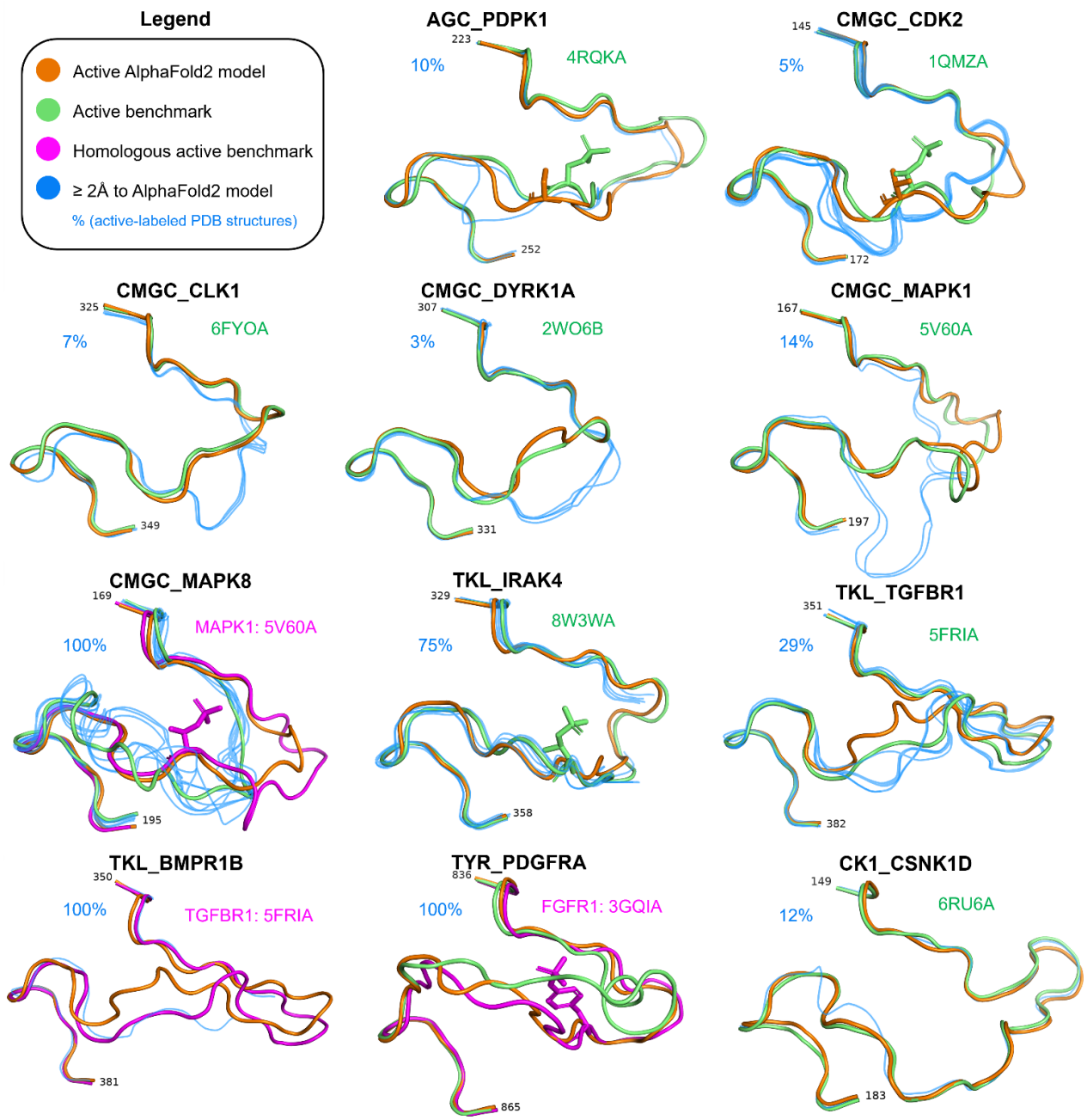
Supplementary Figure 4. Ramachandran distributions for all human kinase domains in the PDB for residues APE6-APE10 and the His and Arg residues of the HRD motif.

Distributions are considerably narrower than the broad ranges defined in the rules for labeling active structures. Each plot shows TYR (top row) and non-TYR (bottom row) kinases separately. The backbone-dihedral criteria are not used for non-TYR kinases in residues APE9 and APE10. For APE6-APE7, the state “APE67-dihe-inflip” indicates accepted structures with a peptide flip (APE7-APE6) conformations of B-L, compared to unflipped structures (“APE67-dihe-in”) which are A-A.



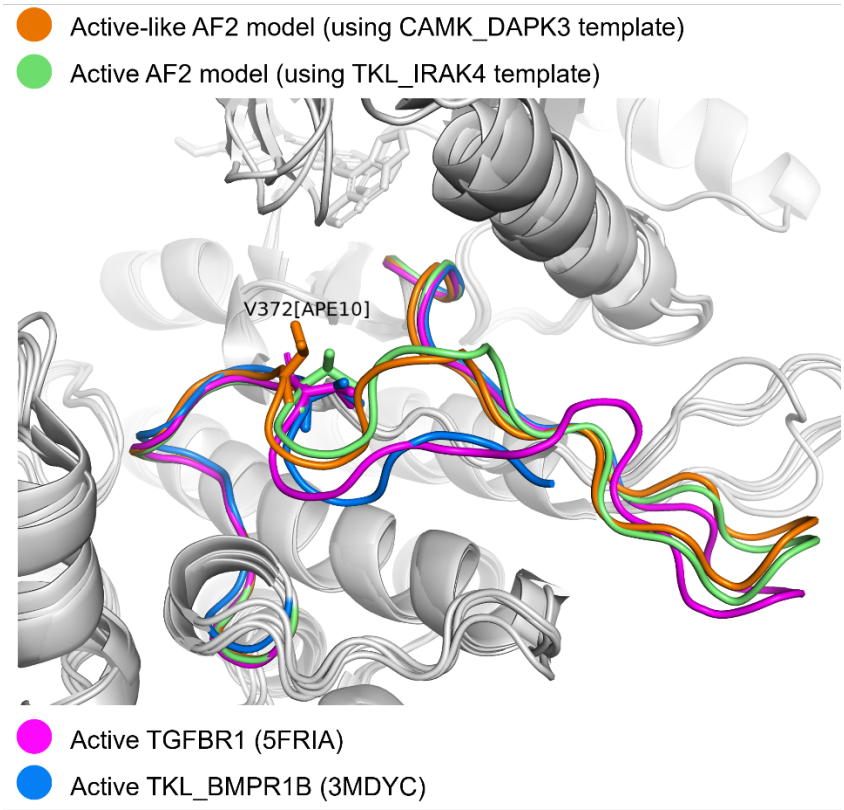
Supplementary Figure 5. Variance and mean of ActLoop ipSAE at different values of d_0.

The variance and mean of ActLoop interaction prediction Scores based on Aligned Errors (ipSAE) were calculated from all 400,000 AlphaFold2 models generated in this study at different values of d_0, with a maximum variance at $d_0 \approx 4$ Å and a mean value of approximately 0.6. The value of 4 Å was used for d_0 in all activation loop ipSAE calculations reported in this study.



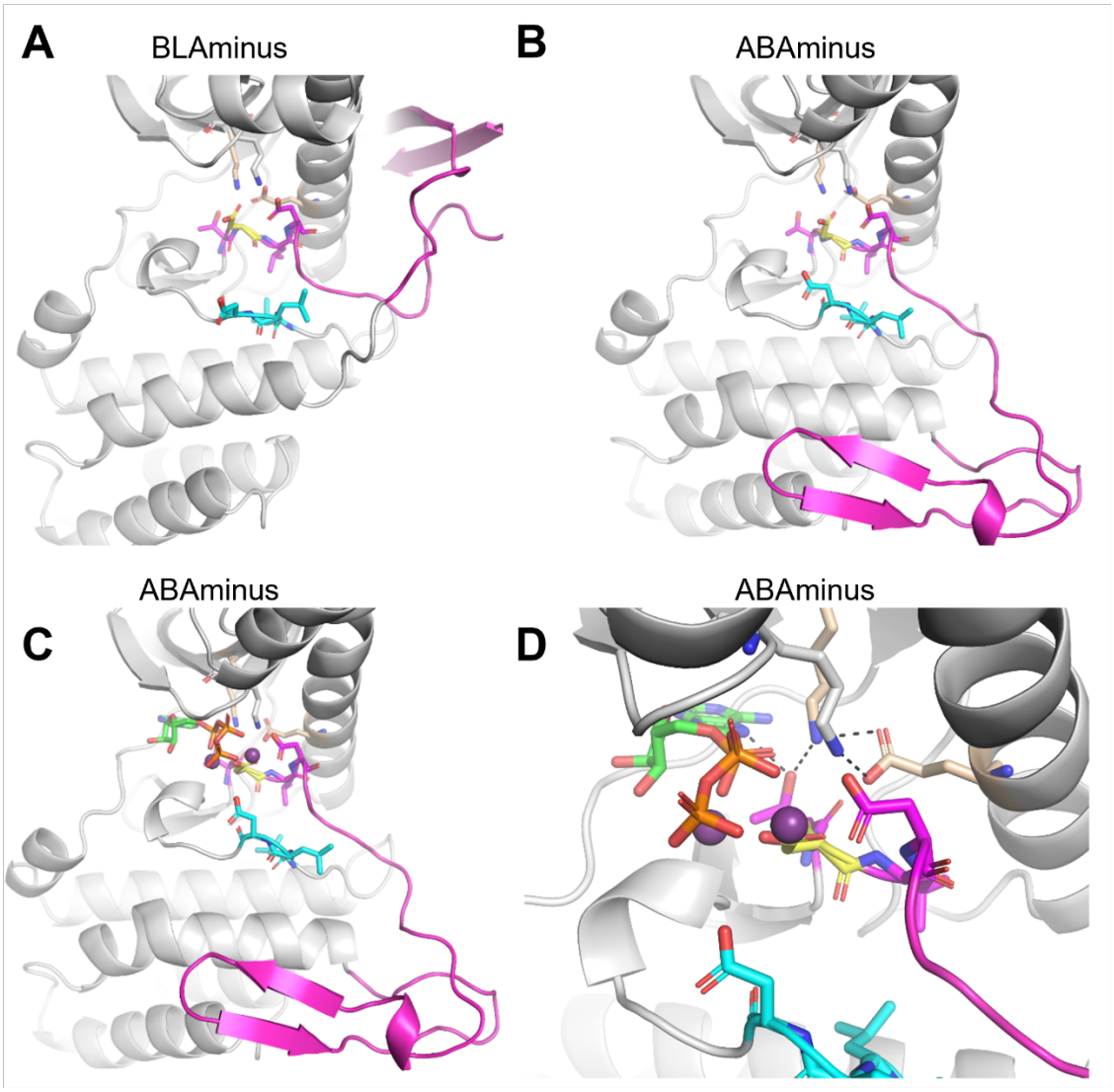
Supplementary Figure 6. Twelve discrepancies between the active AlphaFold2 models and any of their "active" PDB structures.

For each kinase displayed, at least one experimental structure labeled active by the criteria described in this paper can be found which also deviates from the active AlphaFold2 model (orange) in the first 9 or last 15 residues of the activation loop by more than 2 Å. PDB structures which are labeled active but exceed this 2 Å threshold are colored blue, benchmark active structures are colored green and homologous active benchmarks are colored magenta. Also displayed in each panel is the percentage of predicted-active structures which deviate from the active AlphaFold2 model by more than 2 Å. The benchmark structure for CMGC_MAPK8 (green) is likely incorrectly labeled as active – most likely the CMGC_MAPK1 structure 5v60A (pink) is a better active benchmark (see discussion of **Figure 12C** in the main text). For TYR_PDGFRA, the benchmark structure (green) is the only experimental PDB structure with a fully intact activation loop and differs from the active AlphaFold2 model by more than 2 Å (see **Figure 12E** in the main text) – the PDGFRA AlphaFold2 model more closely resembles the phosphorylated/active TYR_FGFR1 benchmark structure 3gqiA (pink).



Supplementary Figure 7. Comparison of TKL_BMPR1B models with related PDB structures.

AlphaFold2 models of TKL_BMPR1B are shown in orange and green. Both models have $ipSAE_{Actloop}$ = 0.77, but the model acquired when using TKL_IRAK4 (PDB: 8W3WA) as a template has a conformation of V372[APE10] that is more compatible with substrate binding.



Supplementary Figure 8. OTHER_PKDCC in a hypothesized active ABAminus conformation.

A. The *BLAminus* conformation generated by AlphaFold2 scores poorly ($ipSAE_{Actloop}$ = 0.27). The sequence of the DFG motif in PKDCC is DLD.

B. The *ABAminus* conformation scores well ($ipSAE_{Actloop}$ =0.84), which we hypothesize to be the active conformation of OTHER_PKDCC.

C. AlphaFold3 models of OTHER_PKDCC bound to ATP·$2Mg^{2+}$ adopt the *ABAminus* conformation.

D. Closeup view of the Alphafold3 model of OTHER_PKDCC and its interactions with ATP in the *ABAminus* conformation. The X of the XDFG (sequence TDLD in PKDCC) and DFG3[Asp] are colored magenta, the β3-αC salt bridge residues are colored tan, the DFG1[Asp] is colored yellow and the HRD (in this case LVD) is colored cyan. An extra lysine in the β2 strand is also shown, which is normally seen in the WNK family and appears to stabilize the DFG and αC-Glu.

Supplementary Table 1. Catalytic kinase domains in the human proteome
The sequence constructs for kinase models produced in this study are given for each human catalytic protein kinase.

N	N (Family)	Family_Gene	SwissProt ID	Gene	Uniprot Acc.	Kinase Start*	Kinase End*	Kinase Length	Protein Length
1	1	AGC_AKT1	AKT1_HUMAN	AKT1	P31749	142	416	275	480
2	2	AGC_AKT2	AKT2_HUMAN	AKT2	P31751	144	417	274	481
3	3	AGC_AKT3	AKT3_HUMAN	AKT3	Q9Y243	140	413	274	479
4	4	AGC_CDC42BPA	MRCKA_HUMAN	CDC42BPA	Q5VT25	69	351	283	1732
5	5	AGC_CDC42BPB	MRCKB_HUMAN	CDC42BPB	Q9Y5S2	68	350	283	1711
6	6	AGC_CDC42BPG	MRCKG_HUMAN	CDC42BPG	Q6DT37	63	345	283	1551
7	7	AGC_CIT	CTRO_HUMAN	CIT	O14578	89	368	280	2027
8	8	AGC_DMPK	DMPK_HUMAN	DMPK	Q09013	63	347	285	629
9	9	AGC_GRK1	GRK1_HUMAN	GRK1	Q15835	182	463	282	563
10	10	AGC_GRK2	ARBK1_HUMAN	GRK2	P25098	183	461	279	689
11	11	AGC_GRK3	ARBK2_HUMAN	GRK3	P35626	183	461	279	688
12	12	AGC_GRK4	GRK4_HUMAN	GRK4	P32298	179	457	279	578
13	13	AGC_GRK5	GRK5_HUMAN	GRK5	P34947	178	456	279	590
14	14	AGC_GRK6	GRK6_HUMAN	GRK6	P43250	178	456	279	576
15	15	AGC_GRK7	GRK7_HUMAN	GRK7	Q8WTQ7	183	462	280	553
16	16	AGC_LATS1	LATS1_HUMAN	LATS1	O95835	697	1018	322	1130
17	17	AGC_LATS2	LATS2_HUMAN	LATS2	Q9NRM7	660	981	322	1088
18	18	AGC_MAST1	MAST1_HUMAN	MAST1	Q9Y2H9	366	655	290	1570
19	19	AGC_MAST2	MAST2_HUMAN	MAST2	Q6P0Q8	504	793	290	1798
20	20	AGC_MAST3	MAST3_HUMAN	MAST3	O60307	359	648	290	1309
21	21	AGC_MAST4	MAST4_HUMAN	MAST4	O15021	562	851	290	2623
22	22	AGC_MASTL	GWL_HUMAN	MASTL	Q96GX5	27	843	817	879
23	23	AGC_PDPK1	PDPK1_HUMAN	PDPK1	O15530	74	350	277	556
24	24	AGC_PKN1	PKN1_HUMAN	PKN1	Q16512	607	882	276	942
25	25	AGC_PKN2	PKN2_HUMAN	PKN2	Q16513	649	924	276	984
26	26	AGC_PKN3	PKN3_HUMAN	PKN3	Q6P5Z2	551	826	276	889
27	27	AGC_PRKACA	KAPCA_HUMAN	PRKACA	P17612	36	306	271	351
28	28	AGC_PRKACB	KAPCB_HUMAN	PRKACB	P22694	36	306	271	351
29	29	AGC_PRKACG	KAPCG_HUMAN	PRKACG	P22612	36	306	271	351
30	30	AGC_PRKCA	KPCA_HUMAN	PRKCA	P17252	331	605	275	672
31	31	AGC_PRKCB	KPCB_HUMAN	PRKCB	P05771	334	608	275	671
32	32	AGC_PRKCD	KPCD_HUMAN	PRKCD	Q05655	341	611	271	676
33	33	AGC_PRKCE	KPCE_HUMAN	PRKCE	Q02156	400	676	277	737
34	34	AGC_PRKCG	KPCG_HUMAN	PRKCG	P05129	343	622	280	697
35	35	AGC_PRKCH	KPCL_HUMAN	PRKCH	P24723	347	622	276	683
36	36	AGC_PRKCI	KPCI_HUMAN	PRKCI	P41743	246	530	285	596
37	37	AGC_PRKCQ	KPCT_HUMAN	PRKCQ	Q04759	372	642	271	706
38	38	AGC_PRKCZ	KPCZ_HUMAN	PRKCZ	Q05513	244	526	283	592
39	39	AGC_PRKG1	KGP1_HUMAN	PRKG1	Q13976	352	627	276	671
40	40	AGC_PRKG2	KGP2_HUMAN	PRKG2	Q13237	445	719	275	762
41	41	AGC_PRKX	PRKX_HUMAN	PRKX	P51817	41	311	271	358
42	42	AGC_ROCK1	ROCK1_HUMAN	ROCK1	Q13464	68	346	279	1354
43	43	AGC_ROCK2	ROCK2_HUMAN	ROCK2	O75116	84	362	279	1388
44	44	AGC_RPS6KA1-1	KS6A1_HUMAN	RPS6KA1	Q15418	54	329	276	735
45	45	AGC_RPS6KA2-1	KS6A2_HUMAN	RPS6KA2	Q15349	51	326	276	733
46	46	AGC_RPS6KA3-1	KS6A3_HUMAN	RPS6KA3	P51812	60	335	276	740
47	47	AGC_RPS6KA4-1	KS6A4_HUMAN	RPS6KA4	O75676	25	309	285	772
48	48	AGC_RPS6KA5-1	KS6A5_HUMAN	RPS6KA5	O75582	41	326	286	802
49	49	AGC_RPS6KA6-1	KS6A6_HUMAN	RPS6KA6	Q9UK32	65	338	274	745
50	50	AGC_RPS6KB1	KS6B1_HUMAN	RPS6KB1	P23443	83	360	278	525
51	51	AGC_RPS6KB2	KS6B2_HUMAN	RPS6KB2	Q9UBS0	59	336	278	482
52	52	AGC_RSKR	KS6R_HUMAN	RSKR	Q96LW2	99	367	269	410
53	53	AGC_SGK1	SGK1_HUMAN	SGK1	O00141	90	363	274	431
54	54	AGC_SGK2	SGK2_HUMAN	SGK2	Q9HBY8	87	360	274	367
55	55	AGC_SGK3	SGK3_HUMAN	SGK3	Q96BR1	154	427	274	496
56	56	AGC_STK32A	ST32A_HUMAN	STK32A	Q8WU08	15	289	275	396
57	57	AGC_STK32B	ST32B_HUMAN	STK32B	Q9NY57	15	291	277	414
58	58	AGC_STK32C	ST32C_HUMAN	STK32C	Q86UX6	85	361	277	486
59	59	AGC_STK38	STK38_HUMAN	STK38	Q15208	81	390	310	465
60	60	AGC_STK38L	ST38L_HUMAN	STK38L	Q9Y2H1	82	391	310	464
61	1	CAMK_AURKA	AURKA_HUMAN	AURKA	O14965	125	391	267	403
62	2	CAMK_AURKB	AURKB_HUMAN	AURKB	Q96GD4	69	335	267	344
63	3	CAMK_AURKC	AURKC_HUMAN	AURKC	Q9UQB9	35	301	267	309
64	4	CAMK_BRSK1	BRSK1_HUMAN	BRSK1	Q8TDC3	26	293	268	778
65	5	CAMK_BRSK2	BRSK2_HUMAN	BRSK2	Q8IWQ3	11	278	268	736
66	6	CAMK_CAMK1	KCC1A_HUMAN	CAMK1	Q14012	12	284	273	370
67	7	CAMK_CAMK1D	KCC1D_HUMAN	CAMK1D	Q8IU85	15	287	273	385
68	8	CAMK_CAMK1G	KCC1G_HUMAN	CAMK1G	Q96NX5	15	285	271	476
69	9	CAMK_CAMK2A	KCC2A_HUMAN	CAMK2A	Q9UQM7	5	279	275	478
70	10	CAMK_CAMK2B	KCC2B_HUMAN	CAMK2B	Q13554	6	280	275	666
71	11	CAMK_CAMK2D	KCC2D_HUMAN	CAMK2D	Q13557	6	280	275	499
72	12	CAMK_CAMK2G	KCC2G_HUMAN	CAMK2G	Q13555	6	280	275	558
73	13	CAMK_CAMK4	KCC4_HUMAN	CAMK4	Q16566	38	308	271	473
74	14	CAMK_CAMKK1	KKCC1_HUMAN	CAMKK1	Q8N5S9	120	417	298	505
75	15	CAMK_CAMKK2	KKCC2_HUMAN	CAMKK2	Q96RR4	157	454	298	588
76	16	CAMK_CHEK1	CHK1_HUMAN	CHEK1	O14757	1	273	273	476
77	17	CAMK_CHEK2	CHK2_HUMAN	CHEK2	O96017	212	494	283	543
78	18	CAMK_DAPK1	DAPK1_HUMAN	DAPK1	P53355	5	283	279	1430
79	19	CAMK_DAPK2	DAPK2_HUMAN	DAPK2	Q9UIK4	15	293	279	370
80	20	CAMK_DAPK3	DAPK3_HUMAN	DAPK3	O43293	5	283	279	454
81	21	CAMK_DCLK1	DCLK1_HUMAN	DCLK1	O15075	382	655	274	740

82	22	CAMK_DCLK2	DCLK2_HUMAN	DCLK2	Q8N568	386	659	274	766
83	23	CAMK_DCLK3	DCLK3_HUMAN	DCLK3	Q9C098	348	621	274	648
84	24	CAMK_HUNK	HUNK_HUMAN	HUNK	P57058	54	328	275	714
85	25	CAMK_KALRN	KALRN_HUMAN	KALRN	O60229	2675	2945	271	2986
86	26	CAMK_MAPKAPK2	MAPK2_HUMAN	MAPKAPK2	P49137	55	333	279	400
87	27	CAMK_MAPKAPK3	MAPK3_HUMAN	MAPKAPK3	Q16644	35	312	278	382
88	28	CAMK_MAPKAPK5	MAPK5_HUMAN	MAPKAPK5	Q8IW41	12	312	301	473
89	29	CAMK_MARK1	MARK1_HUMAN	MARK1	Q9P0L2	52	319	268	795
90	30	CAMK_MARK2	MARK2_HUMAN	MARK2	Q7KZI7	45	312	268	788
91	31	CAMK_MARK3	MARK3_HUMAN	MARK3	P27448	48	315	268	753
92	32	CAMK_MARK4	MARK4_HUMAN	MARK4	Q96L34	51	318	268	752
93	33	CAMK_MELK	MELK_HUMAN	MELK	Q14680	3	271	269	651
94	34	CAMK_MKNK1	MKNK1_HUMAN	MKNK1	Q9BUB5	40	382	343	465
95	35	CAMK_MKNK2	MKNK2_HUMAN	MKNK2	Q9HBH9	75	376	302	465
96	36	CAMK_MYLK	MYLK_HUMAN	MYLK	Q15746	1456	1727	272	1914
97	37	CAMK_MYLK2	MYLK2_HUMAN	MYLK2	Q9H1R3	275	548	274	596
98	38	CAMK_MYLK3	MYLK3_HUMAN	MYLK3	Q32MK0	505	778	274	819
99	39	CAMK_MYLK4	MYLK4_HUMAN	MYLK4	Q86YV6	96	369	274	388
100	40	CAMK_NIM1K	NIM1_HUMAN	NIM1K	Q8IY84	66	333	268	436
101	41	CAMK_NUAK1	NUAK1_HUMAN	NUAK1	O60285	47	314	268	661
102	42	CAMK_NUAK2	NUAK2_HUMAN	NUAK2	Q9H093	45	311	267	628
103	43	CAMK_OBSCN-1	OBSCN_HUMAN	OBSCN	Q5VST9	6460	6729	270	7968
104	44	CAMK_OBSCN-2	OBSCN_HUMAN	OBSCN	Q5VST9	7664	7932	269	7968
105	45	CAMK_PASK	PASK_HUMAN	PASK	Q96RG2	991	1259	269	1323
106	46	CAMK_PHKG1	PHKG1_HUMAN	PHKG1	Q16816	12	296	285	387
107	47	CAMK_PHKG2	PHKG2_HUMAN	PHKG2	P15735	16	299	284	406
108	48	CAMK_PIM1	PIM1_HUMAN	PIM1	P11309	30	298	269	313
109	49	CAMK_PIM2	PIM2_HUMAN	PIM2	Q9P1W9	24	294	271	311
110	50	CAMK_PIM3	PIM3_HUMAN	PIM3	Q86V86	32	301	270	326
111	51	CAMK_PLK1	PLK1_HUMAN	PLK1	P53350	45	313	269	603
112	52	CAMK_PLK2	PLK2_HUMAN	PLK2	Q9NYY3	74	342	269	685
113	53	CAMK_PLK3	PLK3_HUMAN	PLK3	Q9H4B4	54	322	269	646
114	54	CAMK_PLK4	PLK4_HUMAN	PLK4	O00444	4	273	270	970
115	55	CAMK_PNCK	KCC1B_HUMAN	PNCK	Q6P2M8	7	278	272	343
116	56	CAMK_PRKAA1	AAPK1_HUMAN	PRKAA1	Q13131	19	287	269	559
117	57	CAMK_PRKAA2	AAPK2_HUMAN	PRKAA2	P54646	8	276	269	552
118	58	CAMK_PRKD1	KPCD1_HUMAN	PRKD1	Q15139	575	847	273	912
119	59	CAMK_PRKD2	KPCD2_HUMAN	PRKD2	Q9BZL6	543	815	273	878
120	60	CAMK_PRKD3	KPCD3_HUMAN	PRKD3	O94806	568	840	273	890
121	61	CAMK_PSKH1	KPSH1_HUMAN	PSKH1	P11801	90	363	274	424
122	62	CAMK_RPS6KA1-2	KS6A1_HUMAN	RPS6KA1	Q15418	410	683	274	735
123	63	CAMK_RPS6KA2-2	KS6A2_HUMAN	RPS6KA2	Q15349	407	680	274	733
124	64	CAMK_RPS6KA3-2	KS6A3_HUMAN	RPS6KA3	P51812	414	687	274	740
125	65	CAMK_RPS6KA4-2	KS6A4_HUMAN	RPS6KA4	O75676	400	682	283	772
126	66	CAMK_RPS6KA5-2	KS6A5_HUMAN	RPS6KA5	O75582	415	695	281	802
127	67	CAMK_RPS6KA6-2	KS6A6_HUMAN	RPS6KA6	Q9UK32	418	691	274	745
128	68	CAMK_SIK1	SIK1_HUMAN	SIK1	P57059	19	286	268	783
129	69	CAMK_SIK2	SIK2_HUMAN	SIK2	Q9H0K1	12	279	268	926
130	70	CAMK_SIK3	SIK3_HUMAN	SIK3	Q9Y2K2	58	325	268	1321
131	71	CAMK_SNRK	SNRK_HUMAN	SNRK	Q9NRH2	8	277	270	765
132	72	CAMK_SPEG-1	SPEG_HUMAN	SPEG	Q15772	1593	1862	270	3267
133	73	CAMK_SPEG-2	SPEG_HUMAN	SPEG	Q15772	2958	3226	269	3267
134	74	CAMK_STK11	STK11_HUMAN	STK11	Q15831	41	317	277	433
135	75	CAMK_STK17A	ST17A_HUMAN	STK17A	Q9UEE5	51	329	279	414
136	76	CAMK_STK17B	ST17B_HUMAN	STK17B	O94768	24	301	278	372
137	77	CAMK_STK33	STK33_HUMAN	STK33	Q9BYT3	108	389	282	514
138	78	CAMK_TRIO	TRIO_HUMAN	TRIO	O75962	2788	3058	271	3097
139	79	CAMK_TSSK1B	TSSK1_HUMAN	TSSK1B	Q9BXA7	4	280	277	367
140	80	CAMK_TSSK2	TSSK2_HUMAN	TSSK2	Q96PF2	4	280	277	358
141	81	CAMK_TSSK3	TSSK3_HUMAN	TSSK3	Q96PN8	2	268	267	268
142	82	CAMK_TSSK4	TSSK4_HUMAN	TSSK4	Q6SA08	17	301	285	328
143	83	CAMK_TSSK6	TSSK6_HUMAN	TSSK6	Q9BXA6	4	273	270	273
144	1	CK1_CSNK1A1	KC1A_HUMAN	CSNK1A1	P48729	9	291	283	337
145	2	CK1_CSNK1A1L	KC1AL_HUMAN	CSNK1A1L	Q8N752	9	291	283	337
146	3	CK1_CSNK1D	KC1D_HUMAN	CSNK1D	P48730	1	283	283	415
147	4	CK1_CSNK1E	KC1E_HUMAN	CSNK1E	P49674	1	283	283	416
148	5	CK1_CSNK1G1	KC1G1_HUMAN	CSNK1G1	Q9HCP0	36	321	286	422
149	6	CK1_CSNK1G2	KC1G2_HUMAN	CSNK1G2	P78368	38	322	285	415
150	7	CK1_CSNK1G3	KC1G3_HUMAN	CSNK1G3	Q9Y6M4	35	318	284	447
151	8	CK1_TTBK1	TTBK1_HUMAN	TTBK1	Q5TCY1	26	303	278	1321
152	9	CK1_TTBK2	TTBK2_HUMAN	TTBK2	Q6IQ55	13	290	278	1244
153	10	CK1_VRK1	VRK1_HUMAN	VRK1	Q99986	29	334	306	396
154	11	CK1_VRK2	VRK2_HUMAN	VRK2	Q86Y07	21	323	303	508
155	1	CMGC_CDK1	CDK1_HUMAN	CDK1	P06493	1	295	295	297
156	2	CMGC_CDK10	CDK10_HUMAN	CDK10	Q15131	31	331	301	360
157	3	CMGC_CDK11A	CD11A_HUMAN	CDK11A	Q9UQ88	418	719	302	783
158	4	CMGC_CDK11B	CD11B_HUMAN	CDK11B	P21127	430	731	302	795
159	5	CMGC_CDK12	CDK12_HUMAN	CDK12	Q9NYV4	719	1028	310	1490
160	6	CMGC_CDK13	CDK13_HUMAN	CDK13	Q14004	697	1006	310	1512
161	7	CMGC_CDK14	CDK14_HUMAN	CDK14	O94921	127	427	301	469
162	8	CMGC_CDK15	CDK15_HUMAN	CDK15	Q96Q40	95	395	301	435
163	9	CMGC_CDK16	CDK16_HUMAN	CDK16	Q00536	157	454	298	496
164	10	CMGC_CDK17	CDK17_HUMAN	CDK17	Q00537	184	481	298	523
165	11	CMGC_CDK18	CDK18_HUMAN	CDK18	Q07002	136	433	298	474
166	12	CMGC_CDK19	CDK19_HUMAN	CDK19	Q9BWU1	12	343	332	502
167	13	CMGC_CDK2	CDK2_HUMAN	CDK2	P24941	1	294	294	298
168	14	CMGC_CDK20	CDK20_HUMAN	CDK20	Q8IZL9	1	296	296	346

169	15	CMGC_CDK3	CDK3_HUMAN	CDK3	Q00526	1	294	294	305
170	16	CMGC_CDK4	CDK4_HUMAN	CDK4	P11802	1	303	303	303
171	17	CMGC_CDK5	CDK5_HUMAN	CDK5	Q00535	1	292	292	292
172	18	CMGC_CDK6	CDK6_HUMAN	CDK6	Q00534	5	308	304	326
173	19	CMGC_CDK7	CDK7_HUMAN	CDK7	P50613	4	304	301	346
174	20	CMGC_CDK8	CDK8_HUMAN	CDK8	P49336	12	343	332	464
175	21	CMGC_CDK9	CDK9_HUMAN	CDK9	P50750	11	323	313	372
176	22	CMGC_CDKL1	CDKL1_HUMAN	CDKL1	Q00532	1	296	296	358
177	23	CMGC_CDKL2	CDKL2_HUMAN	CDKL2	Q92772	1	295	295	493
178	24	CMGC_CDKL3	CDKL3_HUMAN	CDKL3	Q8IVW4	1	294	294	592
179	25	CMGC_CDKL4	CDKL4_HUMAN	CDKL4	Q5MAI5	1	294	294	379
180	26	CMGC_CDKL5	CDKL5_HUMAN	CDKL5	O76039	5	305	301	960
181	27	CMGC_CLK1	CLK1_HUMAN	CLK1	P49759	153	484	332	484
182	28	CMGC_CLK2	CLK2_HUMAN	CLK2	P49760	155	487	333	499
183	29	CMGC_CLK3	CLK3_HUMAN	CLK3	P49761	296	628	333	638
184	30	CMGC_CLK4	CLK4_HUMAN	CLK4	Q9HAZ1	151	481	331	481
185	31	CMGC_CSNK2A1	CSK21_HUMAN	CSNK2A1	P68400	31	332	302	391
186	32	CMGC_CSNK2A2	CSK22_HUMAN	CSNK2A2	P19784	32	333	302	350
187	33	CMGC_CSNK2A3	CSK23_HUMAN	CSNK2A3	Q8NEV1	31	332	302	391
188	34	CMGC_DYRK1A	DYR1A_HUMAN	DYRK1A	Q13627	151	487	337	763
189	35	CMGC_DYRK1B	DYR1B_HUMAN	DYRK1B	Q9Y463	103	439	337	629
190	36	CMGC_DYRK2	DYRK2_HUMAN	DYRK2	Q92630	214	543	330	601
191	37	CMGC_DYRK3	DYRK3_HUMAN	DYRK3	O43781	201	530	330	588
192	38	CMGC_DYRK4	DYRK4_HUMAN	DYRK4	Q9NR20	96	408	313	520
193	39	CMGC_GSK3A	GSK3A_HUMAN	GSK3A	P49840	111	411	301	483
194	40	CMGC_GSK3B	GSK3B_HUMAN	GSK3B	P49841	48	348	301	420
195	41	CMGC_HIPK1	HIPK1_HUMAN	HIPK1	Q86Z02	182	526	345	1210
196	42	CMGC_HIPK2	HIPK2_HUMAN	HIPK2	Q9H2X6	191	535	345	1198
197	43	CMGC_HIPK3	HIPK3_HUMAN	HIPK3	Q9H422	189	533	345	1215
198	44	CMGC_HIPK4	HIPK4_HUMAN	HIPK4	Q8NE63	3	355	353	616
199	45	CMGC_ICK	CILK1_HUMAN	CILK1	Q9UPZ9	1	292	292	632
200	46	CMGC_MAK	MAK_HUMAN	MAK	P20794	1	292	292	623
201	47	CMGC_MAPK1	MK01_HUMAN	MAPK1	P28482	17	321	305	360
202	48	CMGC_MAPK10	MK10_HUMAN	MAPK10	P53779	56	367	312	464
203	49	CMGC_MAPK11	MK11_HUMAN	MAPK11	Q15759	16	316	301	364
204	50	CMGC_MAPK12	MK12_HUMAN	MAPK12	P53778	19	319	301	367
205	51	CMGC_MAPK13	MK13_HUMAN	MAPK13	O15264	17	316	300	365
206	52	CMGC_MAPK14	MK14_HUMAN	MAPK14	Q16539	16	316	301	360
207	53	CMGC_MAPK15	MK15_HUMAN	MAPK15	Q8TD08	5	312	308	544
208	54	CMGC_MAPK3	MK03_HUMAN	MAPK3	P27361	34	338	305	379
209	55	CMGC_MAPK4	MK04_HUMAN	MAPK4	P31152	12	320	309	587
210	56	CMGC_MAPK6	MK06_HUMAN	MAPK6	Q16659	12	324	313	721
211	57	CMGC_MAPK7	MK07_HUMAN	MAPK7	Q13164	47	355	309	816
212	58	CMGC_MAPK8	MK08_HUMAN	MAPK8	P45983	18	329	312	427
213	59	CMGC_MAPK9	MK09_HUMAN	MAPK9	P45984	18	329	312	424
214	60	CMGC_MOK	MOK_HUMAN	MOK	Q9UQ07	1	293	293	419
215	61	CMGC_NLK	NLK_HUMAN	NLK	Q9UBE8	130	435	306	527
216	62	CMGC_PRP4K	PRP4K_HUMAN	PRP4K	Q13523	679	1007	329	1007
217	63	CMGC_SRPK1	SRPK1_HUMAN	SRPK1	Q96SB4	72	655	584	655
218	64	CMGC_SRPK2	SRPK2_HUMAN	SRPK2	P78362	73	688	616	688
219	65	CMGC_SRPK3	SRPK3_HUMAN	SRPK3	Q9UPE1	71	567	497	567
220	1	NEK_NEK1	NEK1_HUMAN	NEK1	Q96PY6	1	266	266	1258
221	2	NEK_NEK2	NEK2_HUMAN	NEK2	P51955	1	279	279	445
222	3	NEK_NEK3	NEK3_HUMAN	NEK3	P51956	1	265	265	506
223	4	NEK_NEK4	NEK4_HUMAN	NEK4	P51957	1	269	269	841
224	5	NEK_NEK5	NEK5_HUMAN	NEK5	Q6P3R8	1	267	267	708
225	6	NEK_NEK6	NEK6_HUMAN	NEK6	Q9HC98	37	313	277	313
226	7	NEK_NEK7	NEK7_HUMAN	NEK7	Q8TDX7	26	302	277	302
227	8	NEK_NEK8	NEK8_HUMAN	NEK8	Q86SG6	1	266	266	692
228	9	NEK_NEK9	NEK9_HUMAN	NEK9	Q8TD19	44	316	273	979
229	10	NEK_NEK10	NEK10_HUMAN	NEK10	Q6ZWH5	511	793	283	1172
230	11	NEK_NEK11	NEK11_HUMAN	NEK11	Q8NG66	21	295	275	645
231	1	OTHER_AAK1	AAK1_HUMAN	AAK1	Q2M2I8	38	321	284	961
232	2	OTHER_BMP2K	BMP2K_HUMAN	BMP2K	Q9NSY1	43	325	283	1161
233	3	OTHER_BUB1	BUB1_HUMAN	BUB1	O43683	779	1064	286	1085
234	4	OTHER_CDC7	CDC7_HUMAN	CDC7	O00311	50	574	525	574
235	5	OTHER_CHUK	IKKA_HUMAN	CHUK	O15111	7	317	311	745
236	6	OTHER_DSTYK	DUSTY_HUMAN	DSTYK	Q6XUX3	640	916	277	929
237	7	OTHER_EIF2AK1	E2AK1_HUMAN	EIF2AK1	Q9BQI3	159	591	433	630
238	8	OTHER_EIF2AK2	E2AK2_HUMAN	EIF2AK2	P19525	259	546	288	551
239	9	OTHER_EIF2AK3	E2AK3_HUMAN	EIF2AK3	Q9NZJ5	585	1085	501	1116
240	10	OTHER_EIF2AK4-2	E2AK4_HUMAN	EIF2AK4	Q9P2K8	582	1009	428	1649
241	11	OTHER_ERN1	ERN1_HUMAN	ERN1	O75460	561	840	280	977
242	12	OTHER_ERN2	ERN2_HUMAN	ERN2	Q76MJ5	510	789	280	926
243	13	OTHER_GAK	GAK_HUMAN	GAK	O14976	32	323	292	1311
244	14	OTHER_HASPIN	HASP_HUMAN	GSG2	Q8TF76	476	798	323	798
245	15	OTHER_IKBKB	IKKB_HUMAN	IKBKB	O14920	7	316	310	756
246	16	OTHER_IKBKE	IKKE_HUMAN	IKBKE	Q14164	1	312	312	716
247	17	OTHER_MOS	MOS_HUMAN	MOS	P00540	52	346	295	346
248	18	OTHER_PBK	TOPK_HUMAN	PBK	Q96KB5	24	322	299	322
249	19	OTHER_PDIK1L	PDK1L_HUMAN	PDIK1L	Q8N165	1	339	339	341
250	20	OTHER_PINK1	PINK1_HUMAN	PINK1	Q9BXM7	148	517	370	581
251	21	OTHER_PKDCC	PKDCC_HUMAN	PKDCC	Q504Y2	130	399	270	493
252	22	OTHER_PKMYT1	PMYT1_HUMAN	PKMYT1	Q99640	102	367	266	499
253	23	OTHER_SBK1	SBK1_HUMAN	SBK1	Q52WX2	45	323	279	424
254	24	OTHER_SBK2	SBK2_HUMAN	SBK2	P0C263	54	335	282	348
255	25	OTHER_SBK3	SBK3_HUMAN	SBK3	P0C264	35	314	280	359

256	26	OTHER_STK16	STK16_HUMAN	STK16	O75716	12	300	289	305
257	27	OTHER_STK35	STK35_HUMAN	STK35	Q8TDR2	194	534	341	534
258	28	OTHER_STK36	STK36_HUMAN	STK36	Q9NRP7	1	262	262	1315
259	29	OTHER_TBK1	TBK1_HUMAN	TBK1	Q9UHD2	1	312	312	729
260	30	OTHER_TLK1	TLK1_HUMAN	TLK1	Q9UKI8	448	742	295	766
261	31	OTHER_TLK2	TLK2_HUMAN	TLK2	Q86UE8	454	749	296	772
262	32	OTHER_TP53RK	PRPK_HUMAN	TP53RK	Q96S44	25	253	229	253
263	33	OTHER_TTK	TTK_HUMAN	TTK	P33981	517	799	283	857
264	34	OTHER_UHMK1	UHMK1_HUMAN	UHMK1	Q8TAS1	15	312	298	419
265	35	OTHER_ULK1	ULK1_HUMAN	ULK1	O75385	6	286	281	1050
266	36	OTHER_ULK2	ULK2_HUMAN	ULK2	Q8IYT8	1	279	279	1036
267	37	OTHER_ULK3	ULK3_HUMAN	ULK3	Q6PHR2	6	278	273	472
268	38	OTHER_WEE1	WEE1_HUMAN	WEE1	P30291	291	577	287	646
269	39	OTHER_WEE2	WEE2_HUMAN	WEE2	P0C1S8	204	494	291	567
270	40	OTHER_WNK1	WNK1_HUMAN	WNK1	Q9H4A3	212	487	276	2382
271	41	OTHER_WNK2	WNK2_HUMAN	WNK2	Q9Y3S1	186	461	276	2297
272	42	OTHER_WNK3	WNK3_HUMAN	WNK3	Q9BYP7	138	413	276	1800
273	43	OTHER_WNK4	WNK4_HUMAN	WNK4	Q96J92	165	440	276	1243
274	1	STE_MAP2K1	MP2K1_HUMAN	MAP2K1	Q02750	60	369	310	393
275	2	STE_MAP2K2	MP2K2_HUMAN	MAP2K2	P36507	64	377	314	400
276	3	STE_MAP2K3	MP2K3_HUMAN	MAP2K3	P46734	56	333	278	347
277	4	STE_MAP2K4	MP2K4_HUMAN	MAP2K4	P45985	94	375	282	399
278	5	STE_MAP2K5	MP2K5_HUMAN	MAP2K5	Q13163	158	427	270	448
279	6	STE_MAP2K6	MP2K6_HUMAN	MAP2K6	P52564	45	322	278	334
280	7	STE_MAP2K7	MP2K7_HUMAN	MAP2K7	O14733	112	388	277	419
281	8	STE_MAP3K1	M3K1_HUMAN	MAP3K1	Q13233	1235	1512	278	1512
282	9	STE_MAP3K14	M3K14_HUMAN	MAP3K14	Q99558	391	661	271	947
283	10	STE_MAP3K15	M3K15_HUMAN	MAP3K15	Q6ZN16	637	916	280	1313
284	11	STE_MAP3K19	M3K19_HUMAN	MAP3K19	Q56UN5	1053	1328	276	1328
285	12	STE_MAP3K2	M3K2_HUMAN	MAP3K2	Q9Y2U5	348	619	272	619
286	13	STE_MAP3K3	M3K3_HUMAN	MAP3K3	Q99759	354	626	273	626
287	14	STE_MAP3K4	M3K4_HUMAN	MAP3K4	Q9Y6R4	1335	1608	274	1608
288	15	STE_MAP3K5	M3K5_HUMAN	MAP3K5	Q99683	665	946	282	1374
289	16	STE_MAP3K6	M3K6_HUMAN	MAP3K6	O95382	635	914	280	1288
290	17	STE_MAP3K8	M3K8_HUMAN	MAP3K8	P41279	119	396	278	467
291	18	STE_MAP4K1	M4K1_HUMAN	MAP4K1	Q92918	9	282	274	833
292	19	STE_MAP4K2	M4K2_HUMAN	MAP4K2	Q12851	8	281	274	820
293	20	STE_MAP4K3	M4K3_HUMAN	MAP4K3	Q8IVH8	8	281	274	894
294	21	STE_MAP4K4	M4K4_HUMAN	MAP4K4	O95819	17	297	281	1239
295	22	STE_MAP4K5	M4K5_HUMAN	MAP4K5	Q9Y4K4	12	285	274	846
296	23	STE_MINK1	MINK1_HUMAN	MINK1	Q8N4C8	17	297	281	1332
297	24	STE_MYO3A	MYO3A_HUMAN	MYO3A	Q8NEV4	13	295	283	1616
298	25	STE_MYO3B	MYO3B_HUMAN	MYO3B	Q8WXR4	19	301	283	1341
299	26	STE_NRK	NRK_HUMAN	NRK	Q7Z2Y5	17	321	305	1582
300	27	STE_OXSR1	OXSR1_HUMAN	OXSR1	O95747	9	299	291	527
301	28	STE_PAK1	PAK1_HUMAN	PAK1	Q13153	262	529	268	545
302	29	STE_PAK2	PAK2_HUMAN	PAK2	Q13177	241	507	267	524
303	30	STE_PAK3	PAK3_HUMAN	PAK3	O75914	275	542	268	559
304	31	STE_PAK4	PAK4_HUMAN	PAK4	O96013	313	580	268	591
305	32	STE_PAK5	PAK5_HUMAN	PAK5	Q9P286	441	708	268	719
306	33	STE_PAK6	PAK6_HUMAN	PAK6	Q9NQU5	399	666	268	681
307	34	STE_SLK	SLK_HUMAN	SLK	Q9H2G2	26	300	275	1235
308	35	STE_STK10	STK10_HUMAN	STK10	O94804	28	302	275	968
309	36	STE_STK24	STK24_HUMAN	STK24	Q9Y6E0	28	294	267	443
310	37	STE_STK25	STK25_HUMAN	STK25	O00506	12	278	267	426
311	38	STE_STK26	STK26_HUMAN	STK26	Q9P289	16	282	267	416
312	39	STE_STK3	STK3_HUMAN	STK3	Q13188	19	286	268	491
313	40	STE_STK39	STK39_HUMAN	STK39	Q9UEW8	55	345	291	545
314	41	STE_STK4	STK4_HUMAN	STK4	Q13043	22	289	268	487
315	42	STE_TAOK1	TAOK1_HUMAN	TAOK1	Q7L7X3	20	289	270	1001
316	43	STE_TAOK2	TAOK2_HUMAN	TAOK2	Q9UL54	20	289	270	1235
317	44	STE_TAOK3	TAOK3_HUMAN	TAOK3	Q9H2K8	16	285	270	898
318	45	STE_TNIK	TNIK_HUMAN	TNIK	Q9UKE5	17	297	281	1360
319	1	TKL_ACVR1	ACVR1_HUMAN	ACVR1	Q04771	200	505	306	509
320	2	TKL_ACVR1B	ACV1B_HUMAN	ACVR1B	P36896	199	504	306	505
321	3	TKL_ACVR1C	ACV1C_HUMAN	ACVR1C	Q8NER5	187	492	306	493
322	4	TKL_ACVR2A	AVR2A_HUMAN	ACVR2A	P27037	184	489	306	513
323	5	TKL_ACVR2B	AVR2B_HUMAN	ACVR2B	Q13705	182	488	307	512
324	6	TKL_ACVRL1	ACVL1_HUMAN	ACVRL1	P37023	194	499	306	503
325	7	TKL_AMHR2	AMHR2_HUMAN	AMHR2	Q16671	195	515	321	573
326	8	TKL_ANKK1	ANKK1_HUMAN	ANKK1	Q8NFD2	13	295	283	765
327	9	TKL_ARAF	ARAF_HUMAN	ARAF	P10398	302	577	276	606
328	10	TKL_BMPR1A	BMR1A_HUMAN	BMPR1A	P36894	226	531	306	532
329	11	TKL_BMPR1B	BMR1B_HUMAN	BMPR1B	O00238	196	501	306	502
330	12	TKL_BMPR2	BMPR2_HUMAN	BMPR2	Q13873	195	511	317	1038
331	13	TKL_BRAF	BRAF_HUMAN	BRAF	P15056	449	724	276	766
332	14	TKL_IRAK1	IRAK1_HUMAN	IRAK1	P51617	199	529	331	712
333	15	TKL_IRAK4	IRAK4_HUMAN	IRAK4	Q9NWZ3	167	460	294	460
334	16	TKL_LIMK1	LIMK1_HUMAN	LIMK1	P53667	331	614	284	647
335	17	TKL_LIMK2	LIMK2_HUMAN	LIMK2	P53671	323	611	289	638
336	18	TKL_LRRK1	LRRK1_HUMAN	LRRK1	Q38SD2	1230	1530	301	2015
337	19	TKL_LRRK2	LRRK2_HUMAN	LRRK2	Q5S007	1867	2142	276	2527
338	20	TKL_MAP3K10	M3K10_HUMAN	MAP3K10	Q02779	90	367	278	954
339	21	TKL_MAP3K11	M3K11_HUMAN	MAP3K11	Q16584	109	386	278	847
340	22	TKL_MAP3K12	M3K12_HUMAN	MAP3K12	Q12852	117	374	258	859
341	23	TKL_MAP3K13	M3K13_HUMAN	MAP3K13	O43283	160	417	258	966
342	24	TKL_MAP3K20	M3K20_HUMAN	MAP3K20	Q9NYL2	8	270	263	800

343	25	TKL_MAP3K21	M3K21_HUMAN	MAP3K21	Q5TCX8	116	408	293	1036
344	26	TKL_MAP3K7	M3K7_HUMAN	MAP3K7	O43318	28	294	267	606
345	27	TKL_MAP3K9	M3K9_HUMAN	MAP3K9	P80192	136	413	278	1104
346	28	TKL_RAF1	RAF1_HUMAN	RAF1	P04049	341	616	276	648
347	29	TKL_RIPK1	RIPK1_HUMAN	RIPK1	Q13546	9	295	287	671
348	30	TKL_RIPK2	RIPK2_HUMAN	RIPK2	O43353	10	300	291	540
349	31	TKL_RIPK3	RIPK3_HUMAN	RIPK3	Q9Y572	13	293	281	518
350	32	TKL_RIPK4	RIPK4_HUMAN	RIPK4	P57078	14	293	280	832
351	33	TKL_TESK1	TESK1_HUMAN	TESK1	Q15569	46	321	276	626
352	34	TKL_TESK2	TESK2_HUMAN	TESK2	Q96S53	48	319	272	571
353	35	TKL_TGFBR1	TGFR1_HUMAN	TGFBR1	P36897	197	502	306	503
354	36	TKL_TGFBR2	TGFR2_HUMAN	TGFBR2	P37173	236	548	313	567
355	37	TKL_TNNI3K	TNI3K_HUMAN	TNNI3K	Q59H18	455	729	275	835
356	1	TYR_AATK	LMTK1_HUMAN	AATK	Q6ZMQ8	117	405	289	1374
357	2	TYR_ABL1	ABL1_HUMAN	ABL1	P00519	234	503	270	1130
358	3	TYR_ABL2	ABL2_HUMAN	ABL2	P42684	280	549	270	1182
359	4	TYR_ALK	ALK_HUMAN	ALK	Q9UM73	1108	1393	286	1620
360	5	TYR_AXL	UFO_HUMAN	AXL	P30530	528	813	286	894
361	6	TYR_BLK	BLK_HUMAN	BLK	P51451	233	500	268	505
362	7	TYR_BMX	BMX_HUMAN	BMX	P51813	409	675	267	675
363	8	TYR_BTK	BTK_HUMAN	BTK	Q06187	394	659	266	659
364	9	TYR_CSF1R	CSF1R_HUMAN	CSF1R	P07333	574	920	347	972
365	10	TYR_CSK	CSK_HUMAN	CSK	P41240	187	450	264	450
366	11	TYR_DDR1	DDR1_HUMAN	DDR1	Q08345	602	913	312	913
367	12	TYR_DDR2	DDR2_HUMAN	DDR2	Q16832	555	855	301	855
368	13	TYR_EGFR	EGFR_HUMAN	EGFR	P00533	704	978	275	1210
369	14	TYR_EPHA1	EPHA1_HUMAN	EPHA1	P21709	616	890	275	976
370	15	TYR_EPHA2	EPHA2_HUMAN	EPHA2	P29317	605	881	277	976
371	16	TYR_EPHA3	EPHA3_HUMAN	EPHA3	P29320	613	888	276	983
372	17	TYR_EPHA4	EPHA4_HUMAN	EPHA4	P54764	613	888	276	986
373	18	TYR_EPHA5	EPHA5_HUMAN	EPHA5	P54756	667	942	276	1037
374	19	TYR_EPHA6	EPHA6_HUMAN	EPHA6	Q9UF33	623	940	318	1036
375	20	TYR_EPHA7	EPHA7_HUMAN	EPHA7	Q15375	625	900	276	998
376	21	TYR_EPHA8	EPHA8_HUMAN	EPHA8	P29322	627	902	276	1005
377	22	TYR_EPHB1	EPHB1_HUMAN	EPHB1	P54762	611	888	278	984
378	23	TYR_EPHB2	EPHB2_HUMAN	EPHB2	P29323	613	890	278	1055
379	24	TYR_EPHB3	EPHB3_HUMAN	EPHB3	P54753	625	902	278	998
380	25	TYR_EPHB4	EPHB4_HUMAN	EPHB4	P54760	607	884	278	987
381	26	TYR_ERBB2	ERBB2_HUMAN	ERBB2	P04626	712	986	275	1255
382	27	TYR_ERBB4	ERBB4_HUMAN	ERBB4	Q15303	710	984	275	1308
383	28	TYR_FER	FER_HUMAN	FER	P16591	555	822	268	822
384	29	TYR_FES	FES_HUMAN	FES	P07332	553	822	270	822
385	30	TYR_FGFR1	FGFR1_HUMAN	FGFR1	P11362	470	764	295	822
386	31	TYR_FGFR2	FGFR2_HUMAN	FGFR2	P21802	473	767	295	821
387	32	TYR_FGFR3	FGFR3_HUMAN	FGFR3	P22607	464	758	295	806
388	33	TYR_FGFR4	FGFR4_HUMAN	FGFR4	P22455	459	753	295	802
389	34	TYR_FGR	FGR_HUMAN	FGR	P09769	255	522	268	529
390	35	TYR_FLT1	VGFR1_HUMAN	FLT1	P17948	819	1164	346	1338
391	36	TYR_FLT3	FLT3_HUMAN	FLT3	P36888	602	953	352	993
392	37	TYR_FLT4	VGFR3_HUMAN	FLT4	P35916	837	1179	343	1363
393	38	TYR_FRK	FRK_HUMAN	FRK	P42685	226	497	272	505
394	39	TYR_FYN	FYN_HUMAN	FYN	P06241	263	530	268	537
395	40	TYR_HCK	HCK_HUMAN	HCK	P08631	254	521	268	526
396	41	TYR_IGF1R	IGF1R_HUMAN	IGF1R	P08069	991	1276	286	1367
397	42	TYR_INSR	INSR_HUMAN	INSR	P06213	1015	1300	286	1382
398	43	TYR_INSRR	INSRR_HUMAN	INSRR	P14616	971	1256	286	1297
399	44	TYR_ITK	ITK_HUMAN	ITK	Q08881	355	620	266	620
400	45	TYR_JAK1-2	JAK1_HUMAN	JAK1	P23458	867	1154	288	1154
401	46	TYR_JAK2-2	JAK2_HUMAN	JAK2	O60674	841	1132	292	1132
402	47	TYR_JAK3-2	JAK3_HUMAN	JAK3	P52333	814	1105	292	1124
403	48	TYR_KDR	VGFR2_HUMAN	KDR	P35968	826	1170	345	1356
404	49	TYR_KIT	KIT_HUMAN	KIT	P10721	581	934	354	976
405	50	TYR_LCK	LCK_HUMAN	LCK	P06239	237	504	268	509
406	51	TYR_LMTK2	LMTK2_HUMAN	LMTK2	Q8IWU2	129	417	289	1503
407	52	TYR_LMTK3	LMTK3_HUMAN	LMTK3	Q96Q04	125	418	294	1460
408	53	TYR_LTK	LTK_HUMAN	LTK	P29376	502	787	286	864
409	54	TYR_LYN	LYN_HUMAN	LYN	P07948	239	507	269	512
410	55	TYR_MATK	MATK_HUMAN	MATK	P42679	227	488	262	507
411	56	TYR_MERTK	MERTK_HUMAN	MERTK	Q12866	579	864	286	999
412	57	TYR_MET	MET_HUMAN	MET	P08581	1068	1347	280	1390
413	58	TYR_MST1R	RON_HUMAN	MST1R	Q04912	1072	1351	280	1400
414	59	TYR_MUSK	MUSK_HUMAN	MUSK	O15146	567	866	300	869
415	60	TYR_NTRK1	NTRK1_HUMAN	NTRK1	P04629	502	791	290	796
416	61	TYR_NTRK2	NTRK2_HUMAN	NTRK2	Q16620	530	817	288	822
417	62	TYR_NTRK3	NTRK3_HUMAN	NTRK3	Q16288	530	834	305	839
418	63	TYR_PDGFRA	PGFRA_HUMAN	PDGFRA	P16234	585	960	376	1089
419	64	TYR_PDGFRB	PGFRB_HUMAN	PDGFRB	P09619	592	968	377	1106
420	65	TYR_PTK2	FAK1_HUMAN	PTK2	Q05397	414	686	273	1052
421	66	TYR_PTK2B	FAK2_HUMAN	PTK2B	Q14289	417	689	273	1009
422	67	TYR_PTK6	PTK6_HUMAN	PTK6	Q13882	183	451	269	451
423	68	TYR_RET	RET_HUMAN	RET	P07949	716	1015	300	1114
424	69	TYR_ROS1	ROS1_HUMAN	ROS1	P08922	1937	2225	289	2347
425	71	TYR_SRC	SRC_HUMAN	SRC	P12931	262	529	268	536
426	72	TYR_SRMS	SRMS_HUMAN	SRMS	Q9H3Y6	222	488	267	488
427	73	TYR_SYK	KSYK_HUMAN	SYK	P43405	362	635	274	635
428	74	TYR_TEC	TEC_HUMAN	TEC	P42680	362	629	268	631
429	75	TYR_TEK	TIE2_HUMAN	TEK	Q02763	816	1102	287	1124

430	76	TYR_TIE1	TIE1_HUMAN	TIE1	P35590	831	1117	287	1138
431	77	TYR_TNK1	TNK1_HUMAN	TNK1	Q13470	108	387	280	666
432	78	TYR_TNK2	ACK1_HUMAN	TNK2	Q07912	118	395	278	1038
433	79	TYR_TXK	TXK_HUMAN	TXK	P42681	263	527	265	527
434	80	TYR_TYK2-2	TYK2_HUMAN	TYK2	P29597	889	1179	291	1187
435	81	TYR_TYRO3	TYRO3_HUMAN	TYRO3	Q06418	510	796	287	890
436	82	TYR_YES1	YES_HUMAN	YES1	P07947	269	536	268	543
437	83	TYR_ZAP70	ZAP70_HUMAN	ZAP70	P43403	329	603	275	619

* Constructs are 8 amino acids longer at the N and C terminus compared to the sequence alignment in Ref. 3.

Supplementary Table 2. Pseudokinase domains in the human proteome

N	N (Family)	Family_Gene	SwissProt ID	Gene	Uniprot Acc.	Kinase Start	Kinase End	Kinase Length	Protein Length
1	1	CAMK_CAMKV	CAMKV_HUMAN	CAMKV	Q8NCB2	16	294	279	501
2	2	CAMK_CASK	CSKP_HUMAN	CASK	O14936	4	284	281	926
3	3	CAMK_PSKH2	KPSH2_HUMAN	PSKH2	Q96QS6	55	328	274	385
4	4	CAMK_STK40	STK40_HUMAN	STK40	Q8N2I9	27	338	312	435
5	5	CAMK_TRIB1	TRIB1_HUMAN	TRIB1	Q96RU8	86	346	261	372
6	6	CAMK_TRIB2	TRIB2_HUMAN	TRIB2	Q92519	56	316	261	343
7	7	CAMK_TRIB3	TRIB3_HUMAN	TRIB3	Q96RU7	63	323	261	358
8	8	CAMK_TTN	TITIN_HUMAN	TTN	Q8WZ42	32170	32440	271	34350
9	1	CK1_VRK3	VRK3_HUMAN	VRK3	Q8IV63	158	463	306	474
10	1	OTHER_BUB1B	BUB1B_HUMAN	BUB1B	O60566	758	1029	272	1050
11	2	OTHER_EIF2AK4-1	E2AK4_HUMAN	EIF2AK4	Q9P2K8	272	547	276	1649
12	3	OTHER_MLKL	MLKL_HUMAN	MLKL	Q8NB16	193	471	279	471
13	4	OTHER_NRBP1	NRBP_HUMAN	NRBP1	Q9UHY1	56	335	280	535
14	5	OTHER_NRBP2	NRBP2_HUMAN	NRBP2	Q9NSY0	29	314	286	501
15	6	OTHER_PAN3	PAN3_HUMAN	PAN3	Q58A45	472	759	288	887
16	7	OTHER_PEAK1	PEAK1_HUMAN	PEAK1	Q9H792	1319	1673	355	1746
17	8	OTHER_PEAK3	PEAK3_HUMAN	PEAK3	Q6ZS72	163	405	243	473
18	9	OTHER_PIK3R4	PI3R4_HUMAN	PIK3R4	Q99570	18	321	304	1358
19	10	OTHER_POMK	SG196_HUMAN	POMK	Q9H5K3	73	341	269	350
20	11	OTHER_PRAG1	PRAG1_HUMAN	PRAG1	Q86YV5	984	1335	352	1406
21	12	OTHER_PXK	PXK_HUMAN	PXK	Q7Z7A4	138	404	267	578
22	13	OTHER_RNASEL	RN5A_HUMAN	RNASEL	Q05823	353	594	242	741
23	14	OTHER_RPS6KC1	KS6C1_HUMAN	RPS6KC1	Q96S38	332	1066	735	1066
24	15	OTHER_RPS6KL1	RPKL1_HUMAN	RPS6KL1	Q9Y6S9	145	547	403	549
25	16	OTHER_SCYL1	SCYL1_HUMAN	SCYL1	Q96KG9	6	271	266	808
26	17	OTHER_SCYL2	SCYL2_HUMAN	SCYL2	Q6P3W7	24	335	312	929
27	18	OTHER_SCYL3	PACE1_HUMAN	SCYL3	Q8IZE3	3	253	251	742
28	19	OTHER_STK31	STK31_HUMAN	STK31	Q9BXU1	702	980	279	1019
29	20	OTHER_STKLD1	STKL1_HUMAN	STKLD1	Q8NE28	20	305	286	680
30	21	OTHER_TBCK	TBCK_HUMAN	TBCK	Q8TEA7	1	281	281	893
31	22	OTHER_TEX14	TEX14_HUMAN	TEX14	Q8IWB6	219	520	302	1497
32	23	OTHER_ULK4	ULK4_HUMAN	ULK4	Q96C45	1	288	288	1275
33	1	RGC_GUCY2C	GUC2C_HUMAN	GUCY2C	P25092	468	767	300	1073
34	2	RGC_GUCY2D	GUC2D_HUMAN	GUCY2D	Q02846	502	815	314	1103
35	3	RGC_GUCY2F	GUC2F_HUMAN	GUCY2F	P51841	505	853	349	1108
36	4	RGC_NPR1	ANPRA_HUMAN	NPR1	P16066	507	829	323	1061
37	5	RGC_NPR2	ANPRB_HUMAN	NPR2	P20594	491	814	324	1047
38	1	STE_STRADA	STRAA_HUMAN	STRADA	Q7RTN6	61	387	327	431
39	2	STE_STRADB	STRAB_HUMAN	STRADB	Q9C0K7	50	377	328	418
40	1	TKL_ILK	ILK_HUMAN	ILK	Q13418	185	452	268	452
41	2	TKL_IRAK2	IRAK2_HUMAN	IRAK2	O43187	197	509	313	625
42	3	TKL_IRAK3	IRAK3_HUMAN	IRAK3	Q9Y616	152	453	302	596
43	4	TKL_KSR1	KSR1_HUMAN	KSR1	Q8IVT5	605	887	283	923
44	5	TKL_KSR2	KSR2_HUMAN	KSR2	Q6VAB6	658	938	281	950
45	1	TYR_EPHA10	EPHAA_HUMAN	EPHA10	Q5JZY3	637	910	274	1008
46	2	TYR_EPHB6	EPHB6_HUMAN	EPHB6	O15197	662	925	264	1021
47	3	TYR_ERBB3	ERBB3_HUMAN	ERBB3	P21860	701	975	275	1342
48	4	TYR_JAK1-1	JAK1_HUMAN	JAK1	P23458	575	855	281	1154
49	5	TYR_JAK2-1	JAK2_HUMAN	JAK2	O60674	537	815	279	1132
50	6	TYR_JAK3-1	JAK3_HUMAN	JAK3	P52333	513	787	275	1124
51	7	TYR_PTK7	PTK7_HUMAN	PTK7	Q13308	788	1070	283	1070
52	8	TYR_ROR1	ROR1_HUMAN	ROR1	Q01973	465	756	292	937
53	9	TYR_ROR2	ROR2_HUMAN	ROR2	Q01974	465	756	292	943
54	10	TYR_RYK	RYK_HUMAN	RYK	P34925	322	606	285	607
55	11	TYR_STYK1	STYK1_HUMAN	STYK1	Q6J9G0	105	390	286	422
56	12	TYR_TYK2-1	TYK2_HUMAN	TYK2	P29597	581	876	296	1187

Notes: Pseudokinase proteins may have catalytic activity that does not involve protein phosphorylation. The kinase domain in POMK is a protein O-mannose kinase. RNASEL is a 2'-5' endonuclease, in which the pseudokinase domain facilitates homodimerization. The pseudokinase domain of PAN3 participates in mRNA deadenylation by PAN2. The RGC proteins contain active guanylyl cyclase domains; the kinase domains are inactive. Human PLK5 contains a truncated kinase domain at its N-terminus and is not included in the table. The mouse ortholog contains a full kinase domain.

Supplementary Table 3. Data for 248 substrate-bound kinase structures

	Kinase	PDB	Substrate (Unip.)	Substrate Length	Ligands	XDF	ABAminus EDIA of XDFG-O	# ActLoop missing residues
1	AGC_AKT1_HUMAN	3cquA	GSK3B_HUMAN	10	CQU	BLAminus	-	0
1	AGC_AKT1_HUMAN	3ow4B	GSK3B_HUMAN	10	SMY	BLAminus	-	0
1	AGC_AKT1_HUMAN	3ocbA	GSK3B_HUMAN	10	XM1	BLAminus	-	0
1	AGC_AKT1_HUMAN	3ow4A	GSK3B_HUMAN	10	SMY	BLAminus	-	0
1	AGC_AKT1_HUMAN	3cqwA	GSK3B_HUMAN	10	MN,CQW	BLAminus	-	0
1	AGC_AKT1_HUMAN	3qkkA	GSK3B_HUMAN	10	SMH	BLAminus	-	0
1	AGC_AKT1_HUMAN	3mvhA	GSK3B_HUMAN	10	MN,WFE	BLAminus	-	0
1	AGC_AKT1_HUMAN	3ocbB	GSK3B_HUMAN	10	XM1	BLAminus	-	0
1	AGC_AKT1_HUMAN	3qklA	GSK3B_HUMAN	10	SMR	BLAminus	-	0
1	AGC_AKT1_HUMAN	4ekkB	GSK3B_HUMAN	10	ANP,MN	BLAminus	-	0
1	AGC_AKT1_HUMAN	3mv5A	GSK3B_HUMAN	10	MN,XFE	BLAminus	-	0
1	AGC_AKT1_HUMAN	4ekkA	GSK3B_HUMAN	10	ANP,MN	BLAminus	-	0
2	AGC_AKT2_HUMAN	2uw9A	GSK3B_HUMAN	10	GVP	BLAminus	-	0
2	AGC_AKT2_HUMAN	2x39A	GSK3B_HUMAN	10	X39	BLAminus	-	0
2	AGC_AKT2_HUMAN	2jdoA	GSK3B_HUMAN	10	I5S	BLAminus	-	0
2	AGC_AKT2_HUMAN	2jdrA	GSK3B_HUMAN	10	L20	BLAminus	-	0
2	AGC_AKT2_HUMAN	2xh5A	GSK3B_HUMAN	10	X37	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e88A	GSK3B_HUMAN	10	G96	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e88B	GSK3B_HUMAN	10	G96	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e8dB	GSK3B_HUMAN	10	G98	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e8dA	GSK3B_HUMAN	10	G98	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e87A	GSK3B_HUMAN	10	G95	BLAminus	-	0
2	AGC_AKT2_HUMAN	3e87B	GSK3B_HUMAN	10	G95	BLAminus	-	0
2	AGC_AKT2_HUMAN	1o6lA	GSK3B_HUMAN	10	ANP,MN	BLAminus	-	0
2	AGC_AKT2_HUMAN	1o6kA	GSK3B_HUMAN	10	ANP,MN	BLAminus	-	0
3	AGC_PRKACA_MOUSE	3x2uA	IPKA_HUMAN	20	ATP,MG	BLAminus	-	0
3	AGC_PRKACA_MOUSE	1l3rE	IPKA_HUMAN	20	MG,ADP	BLAminus	-	0
3	AGC_PRKACA_MOUSE	1jluE	IPKA_HUMAN	20	No_ligand	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4dfzE	IPKA_HUMAN	20	No_ligand	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4dg2E	IPKA_HUMAN	20	No_ligand	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4dfxE	IPKA_HUMAN	20	ANP,MG	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4o22A	IPKA_HUMAN	20	No_ligand	BLAminus	-	0
3	AGC_PRKACA_MOUSE	3x2wA	IPKA_HUMAN	20	MG,ATP	BLAminus	-	0
3	AGC_PRKACA_MOUSE	1jbpE	IPKA_HUMAN	20	ADP	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4dg0E	IPKA_HUMAN	20	ANP,MG	BLAminus	-	0
3	AGC_PRKACA_MOUSE	4iacA	IPKA_HUMAN	20	MG,ACP	BLAminus	-	0
4	AGC_PRKACA_MOUSE	4x6qC	KAP3_MOUSE	270	No_ligand	ABAminus	0.75	0
4	AGC_PRKACA_MOUSE	3tnpC	KAP3_MOUSE	270	No_ligand	ABAminus	0.49	0
4	AGC_PRKACA_MOUSE	3tnpF	KAP3_MOUSE	270	No_ligand	ABAminus	0.42	0
4	AGC_PRKACA_MOUSE	3j4qD	KAP3_MOUSE	382	No_ligand	BLAminus	-	0
4	AGC_PRKACA_MOUSE	3j4qE	KAP3_MOUSE	382	No_ligand	BLAminus	-	0
4	AGC_PRKACA_MOUSE	3j4rD	KAP3_MOUSE	382	No_ligand	BLAminus	-	0
4	AGC_PRKACA_MOUSE	3j4rE	KAP3_MOUSE	382	No_ligand	BLAminus	-	0
4	AGC_PRKACA_MOUSE	2qvsE	KAP3_MOUSE	298	No_ligand	BLAminus	-	0
5	AGC_PRKACA_MOUSE	7e12A	PPLA_HUMAN	9	ANP,MG	BLAminus	-	0
6	AGC_PRKACA_MOUSE	6mm7A	RYR2_MOUSE	206	ANP	ABAminus	0.85	0
6	AGC_PRKACA_MOUSE	6mm7D	RYR2_MOUSE	206	ANP	ABAminus	0.91	0
6	AGC_PRKACA_MOUSE	6mm8C	RYR2_MOUSE	206	ANP	ABAminus	0.91	0
6	AGC_PRKACA_MOUSE	6mm6C	RYR2_MOUSE	201	ANP	ABAminus	0.56	0
6	AGC_PRKACA_MOUSE	6mm6E	RYR2_MOUSE	198	ANP	ABAminus	0.63	0
7	AGC_PRKACA_MOUSE	6mm5E	RYR2_MOUSE	12	MG,ANP	BLAminus	-	0
8	AGC_PRKCI_HUMAN	5lihA	KPCE_HUMAN	13	ADP,MN	BLAminus	-	0
8	AGC_PRKCI_HUMAN	5lihB	KPCE_HUMAN	10	ADP,MN	BLAminus	-	0
9	AGC_PRKCI_HUMAN	5li1A	Q28E03_XENTR	20	ANP,MG	BLAminus	-	0
9	AGC_PRKCI_MOUSE	4dc2A	Q28E03_XENTR	16	ADE	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujpA	KCC2A_HUMAN	14	ADP,MG	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujqA	KCC2A_HUMAN	14	UZD	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujqB	KCC2A_HUMAN	12	UZD	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujrA	KCC2A_HUMAN	13	No_ligand	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujsA	KCC2A_HUMAN	16	ADP	BLAminus	-	0
10	CAMK_CAMK2A_HUMAN	7ujpB	KCC2A_HUMAN	12	ATP,MG	BLAminus	-	0
11	CAMK_CAMK2D_HUMAN	2welA	KCC2D_HUMAN	304	K88	BLAminus	-	0
12	CAMK_CAMKII_CAEEL	3kk9A	KCC2D_CAEEL	276	No_ligand	BLAminus	-	0
12	CAMK_CAMKII_CAEEL	3kk8A	KCC2D_CAEEL	284	MG	BLAminus	-	0
13	CAMK_CAMKII_DROME	5h9bA	KCNAE_DROME	17	AN2,MG	BLAminus	-	0
13	CAMK_CAMKII_DROME	5hu3A	KCNAE_DROME	16	ADP,MG	BLAminus	-	0
13	CAMK_CAMKII_DROME	5fg8A	KCNAE_DROME	15	ADP,MG	BLAminus	-	0
14	CAMK_PHKG1_RABIT	2phkA	Peptide	7	MN,ATP	BLAminus	-	0
15	CAMK_PIM1_HUMAN	2bilB	Peptide	7	BI1	BLAminus	-	0
15	CAMK_PIM1_HUMAN	4gw8A	Peptide	9	3RA	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6pdoA	Peptide	8	OD1	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6pdiA	Peptide	10	OCJ	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3ma3A	Peptide	7	01I	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3cxwA	Peptide	10	7CP	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6pdnA	Peptide	8	OD4	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4uA	Peptide	9	8MZ	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3jpvA	Peptide	7	1DR	BLAminus	-	0

15	CAMK_PIM1_HUMAN	8afrA	Peptide	9	M0F	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3qf9A	Peptide	9	NM8	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n5lA	Peptide	9	8NZ	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3cy2A	Peptide	9	MB9	BLAminus	-	0
15	CAMK_PIM1_HUMAN	2bzkB	Peptide	7	ANP	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6qxkB	Peptide	8	JKW	BLAminus	-	0
15	CAMK_PIM1_HUMAN	3cy3A	Peptide	8	JN5	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n50A	Peptide	9	8MN	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6pcwA	Peptide	8	O97	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4yA	Peptide	8	8N5	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5mzlA	Peptide	8	8ET	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n5mA	Peptide	9	8O8	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4zA	Peptide	9	8MK	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n51A	Peptide	8	8N8	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4xA	Peptide	8	8MT	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4vA	Peptide	9	8MW	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4rA	Peptide	9	8MQ	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4nA	Peptide	9	8M8	BLAminus	-	0
15	CAMK_PIM1_HUMAN	2c3iB	Peptide	8	IYZ	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5ndtA	Peptide	9	8UB	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n52A	Peptide	7	8N2	BLAminus	-	0
15	CAMK_PIM1_HUMAN	6pdpA	Peptide	7	OBY	BLAminus	-	0
15	CAMK_PIM1_HUMAN	7qfmA	Peptide	8	AY3	BLAminus	-	0
15	CAMK_PIM1_HUMAN	7z6uA	Peptide	9	IJB	BLAminus	-	0
15	CAMK_PIM1_HUMAN	5n4oA	Peptide	9	8MB	BLAminus	-	0
15	CAMK_PIM1_HUMAN	7qb2A	Peptide	8	A7X	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru6A	P63_HUMAN	8	ACP	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru7A	P63_HUMAN	15	ADP	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru8A	P63_HUMAN	10	ADP	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru8B	P63_HUMAN	12	ADP	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru8D	P63_HUMAN	9	ADP	BLAminus	-	0
16	CK1_CSNK1D_HUMAN	6ru8C	P63_HUMAN	9	ADP	BLAminus	-	6
17	CK1_CSNK1D_HUMAN	8d7oB	PER2_HUMAN	7	No_ligand	BLAminus	-	0
17	CK1_CSNK1D_HUMAN	8d7oA	PER2_HUMAN	7	No_ligand	BLAminus	-	0
17	CK1_CSNK1D_HUMAN	8d7nB	PER2_HUMAN	7	No_ligand	BLAminus	-	0
17	CK1_CSNK1D_HUMAN	8d7mA	PER2_HUMAN	6	No_ligand	BLAminus	-	0
17	CK1_CSNK1D_HUMAN	8d7mB	PER2_HUMAN	6	No_ligand	BLAminus	-	0
17	CK1_CSNK1D_HUMAN	8d7nA	PER2_HUMAN	7	No_ligand	BLAminus	-	0
18	CMGC_CDK2_HUMAN	1qmzA	Peptide	7	ATP,MG	BLAminus	-	0
18	CMGC_CDK2_HUMAN	1qmzC	Peptide	7	ATP,MG	BLAminus	-	0
18	CMGC_CDK2_HUMAN	2cciA	Peptide	19	ATP,MG	BLAminus	-	0
18	CMGC_CDK2_HUMAN	1gy3C	Peptide	7	ATP,MG	BLAminus	-	0
18	CMGC_CDK2_HUMAN	1gy3A	Peptide	7	ATP,MG	BLAminus	-	0
19	CMGC_CDK2_HUMAN	3qhwC	Peptide	10	ADP,MG	BLAminus	-	0
19	CMGC_CDK2_HUMAN	3qhrA	Peptide	10	ADP,MG	BLAminus	-	0
19	CMGC_CDK2_HUMAN	3qhwA	Peptide	10	ADP,MG	BLAminus	-	0
19	CMGC_CDK2_HUMAN	3qhrC	Peptide	10	ADP,MG	BLAminus	-	0
20	CMGC_DYRK1A_HUMAN	2wo6B	Peptide	8	D15	BLAminus	-	0
21	OTHER_CDC7_HUMAN	6ya7A	MCM2_HUMAN	9	ADP	BLAminus	-	123
22	OTHER_HASPIN_HUMAN	4oucA	H32_HUMAN	7	5ID	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4nB	PINK1_PEDHC	404	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4nA	PINK1_PEDHC	399	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mA	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mC	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mE	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mG	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mI	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mK	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mB	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mD	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mF	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mH	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mJ	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4mL	PINK1_PEDHC	408	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	8uyhA	PINK1_PEDHC	370	ANP,MG	BLAminus	-	0
23	OTHER_PINK1_PEDHC	8uyfA	PINK1_PEDHC	401	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	8uyfB	PINK1_PEDHC	400	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	8uyhB	PINK1_PEDHC	375	ANP,MG	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4kB	PINK1_PEDHC	379	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4lA	PINK1_PEDHC	377	No_ligand	BLAminus	-	0
23	OTHER_PINK1_PEDHC	7t4lB	PINK1_PEDHC	378	No_ligand	BLAminus	-	0
23	OTHER_PINK1_TRICA	7mp8A	PINK1_TRICA	399	No_ligand	BLAminus	-	0
23	OTHER_PINK1_TRICA	8uctA	PINK1_TRICA	391	3YT	None	-	0
23	OTHER_PINK1_TRICA	7mp9A	PINK1_TRICA	414	MG,AN2	None	-	0
24	STE_PAK1_HUMAN	6b16A	PAK1_HUMAN	281	C7Y	BLAminus	-	0
24	STE_PAK1_HUMAN	4zloB	PAK1_HUMAN	278	No_ligand	BLAminus	-	0
24	STE_PAK1_HUMAN	4zy4A	PAK1_HUMAN	292	4T3	BLAminus	-	0
24	STE_PAK1_HUMAN	3q4zA	PAK1_HUMAN	279	MG,ANP	BLAminus	-	0
24	STE_PAK1_HUMAN	4zy6A	PAK1_HUMAN	280	4T6	BLAminus	-	0
24	STE_PAK1_HUMAN	4zy5A	PAK1_HUMAN	287	4T5	BLAminus	-	0
25	STE_PAK4_HUMAN	6wlxA	CTNB1_HUMAN	5	No_ligand	BLAminus	-	0
26	STE_PAK4_HUMAN	6wlyA	LIMK1_HUMAN	9	No_ligand	BLAminus	-	0
27	STE_PAK4_HUMAN	4jdjA	Peptide	6	No_ligand	BLAminus	-	0

27	STE_PAK4_HUMAN	4jdkA	Peptide	6	No_ligand	BLAminus	-	0
27	STE_PAK4_HUMAN	4jdiA	Peptide	6	ANP,MG	BLAminus	-	0
27	STE_PAK4_HUMAN	4jdhA	Peptide	9	No_ligand	BLAminus	-	0
28	STE_PAK4_HUMAN	2q0nA	Peptide	11	No_ligand	BLAminus	-	0
29	STE_TNIK_HUMAN	7xzrB	Peptide	18	ANP,MG	BLAminus	-	0
30	TKL_BAK1_ARATH	3tl8G	HPAB2_PSESM	88	No_ligand	None	-	0
30	TKL_BAK1_ARATH	3tl8A	HPAB2_PSESM	87	No_ligand	ABAminus	0.31	0
30	TKL_BAK1_ARATH	3tl8D	HPAB2_PSESM	88	No_ligand	ABAminus	0.55	0
30	TKL_BAK1_ARATH	3tl8H	HPAB2_PSESM	87	No_ligand	ABAminus	0.43	0
31	TKL_IRAK4_HUMAN	4u97A	IRAK4_HUMAN	270	STU	BLAminus	-	8
32	TKL_IRAK4_HUMAN	8wtfA	IRAK4_HUMAN	291	8CG	BLAminus	-	0
33	TYR_ABL1_HUMAN	2g2iB	Peptide	7	ADP	BLAminus	-	0
33	TYR_ABL1_HUMAN	2g2iA	Peptide	8	ADP	BLAminus	-	0
34	TYR_CSF1R_HUMAN	3lcdA	CSF1R_HUMAN	290	BDY	BLAminus	-	2
35	TYR_EGFR_HUMAN	4r3pA	ERRFI_HUMAN	7	No_ligand	BLAminus	-	0
35	TYR_EGFR_HUMAN	4r3rA	ERRFI_HUMAN	7	No_ligand	BLAminus	-	0
36	TYR_EGFR_HUMAN	5czhA	Peptide	9	No_ligand	BLAminus	-	0
37	TYR_EGFR_HUMAN	5cziA	SHC1_HUMAN	7	No_ligand	BLAminus	-	0
38	TYR_EPHA2_HUMAN	4pdoA	EPHA2_HUMAN	261	No_ligand	None	-	18
38	TYR_EPHA2_HUMAN	4pdoB	EPHA2_HUMAN	261	No_ligand	None	-	18
39	TYR_EPHA3_HUMAN	3fxxA	Peptide	9	ANP	BLAminus	-	9
39	TYR_EPHA3_HUMAN	3fy2A	Peptide	8	No_ligand	BLAminus	-	10
40	TYR_FES_HUMAN	3cblA	Peptide	6	STU	BLAminus	-	0
40	TYR_FES_HUMAN	3cd3A	Peptide	6	STU	BLAminus	-	0
41	TYR_FGFR1_HUMAN	3gqiA	FGFR1_HUMAN	307	ACP,MG	BLAminus	-	0
42	TYR_FGFR2_HUMAN	3clyA	FGFR2_HUMAN	291	MG,ACP	BLAminus	-	0
43	TYR_FGFR2_HUMAN	2pvfA	FGFR2_HUMAN	4	MG,ACP	BLAminus	-	0
44	TYR_FGFR2_HUMAN	5eg3A	PLCG1_RAT	113	ACP,MG	BLAminus	-	0
45	TYR_FGFR3_HUMAN	4k33A	FGFR3_HUMAN	293	ACP,MG	BLAminus	-	0
45	TYR_FGFR3_HUMAN	7dhlA	FGFR3_HUMAN	289	H6X	BLAminus	-	0
45	TYR_FGFR3_HUMAN	6lvmA	FGFR3_HUMAN	288	EVR	BLAminus	-	0
46	TYR_FGFR3_HUMAN	8uduB	FGFR3_HUMAN	280	WIQ	BLAminus	-	1
47	TYR_IGF1R_HUMAN	3lvpC	IGF1R_HUMAN	295	PDR	ABAminus	0.54	2
48	TYR_IGF1R_HUMAN	1k3aA	IRS1_HUMAN	8	ACP	ABAminus	0.68	0
49	TYR_INSR_HUMAN	1ir3A	IRS2_MOUSE	6	MG,ANP	BLAminus	-	0
49	TYR_INSR_HUMAN	2z8cA	IRS2_MOUSE	6	S91	BLAminus	-	0
49	TYR_INSR_HUMAN	3bu3A	IRS2_MOUSE	14	No_ligand	BLAminus	-	0
49	TYR_INSR_HUMAN	3bu5A	IRS2_MOUSE	12	MG,ATP	BLAminus	-	0
49	TYR_INSR_HUMAN	3bu6A	IRS2_MOUSE	14	No_ligand	ABAminus	1.01	0
50	TYR_INSR_HUMAN	1rqqA	Peptide	14	MN	BLAminus	-	0
50	TYR_INSR_HUMAN	1rqqB	Peptide	13	MN	BLAminus	-	0
50	TYR_INSR_HUMAN	1gagA	Peptide	13	MG	BLAminus	-	0
51	TYR_JAK2-2_HUMAN	6vncA	JAK2_HUMAN	297	R6V	ABAminus	0.53	0
51	TYR_JAK2-2_HUMAN	6vnkA	JAK2_HUMAN	297	RXT	ABAminus	0.98	0
51	TYR_JAK2-2_HUMAN	6vn8A	JAK2_HUMAN	298	3JW	ABAminus	1.01	0
51	TYR_JAK2-2_HUMAN	6vnmA	JAK2_HUMAN	298	R5Y	ABAminus	0.52	0
51	TYR_JAK2-2_HUMAN	6vglB	JAK2_HUMAN	305	RXT	ABAminus	0.95	0
51	TYR_JAK2-2_HUMAN	6vnbA	JAK2_HUMAN	298	R6P	ABAminus	0.53	0
51	TYR_JAK2-2_HUMAN	8bx9A	JAK2_HUMAN	300	S59	ABAminus	1.04	0
51	TYR_JAK2-2_HUMAN	8bpvA	JAK2_HUMAN	300	6T3	ABAminus	0.5	0
51	TYR_JAK2-2_HUMAN	8bx6A	JAK2_HUMAN	300	S5I	BLAminus	-	0
51	TYR_JAK2-2_HUMAN	6vglA	JAK2_HUMAN	298	RXT	ABAminus	1.02	0
51	TYR_JAK2-2_HUMAN	6vnjA	JAK2_HUMAN	299	R5S	ABAminus	0.8	0
51	TYR_JAK2-2_HUMAN	6vnkB	JAK2_HUMAN	297	RXT	ABAminus	0.89	0
51	TYR_JAK2-2_HUMAN	8bm2B	JAK2_HUMAN	300	QQC	BLAminus	-	0
51	TYR_JAK2-2_HUMAN	6vnhA	JAK2_HUMAN	298	XZS	ABAminus	0.9	0
51	TYR_JAK2-2_HUMAN	6vs3B	JAK2_HUMAN	298	R6V	ABAminus	1.05	0
51	TYR_JAK2-2_HUMAN	8bpwB	JAK2_HUMAN	300	2V9	ABAminus	0.5	0
51	TYR_JAK2-2_HUMAN	6vsnB	JAK2_HUMAN	298	RG4	ABAminus	0.73	0
51	TYR_JAK2-2_HUMAN	8bxhA	JAK2_HUMAN	300	C87	ABAminus	0.82	0
51	TYR_JAK2-2_HUMAN	6vnlB	JAK2_HUMAN	298	5W2	ABAminus	0.69	0
51	TYR_JAK2-2_HUMAN	6vnfA	JAK2_HUMAN	298	R6S	ABAminus	0.92	0
51	TYR_JAK2-2_HUMAN	8bpwA	JAK2_HUMAN	300	2V9	ABAminus	0.56	0
51	TYR_JAK2-2_HUMAN	8bm2A	JAK2_HUMAN	300	QQC	BLAminus	-	0
51	TYR_JAK2-2_HUMAN	6vnlA	JAK2_HUMAN	298	5W2	ABAminus	0.83	0
52	TYR_KIT_HUMAN	7khjA	KIT_HUMAN	288	WEG	BLAminus	-	2
52	TYR_KIT_HUMAN	8pqfC	KIT_HUMAN	297	9KI	BLAminus	-	2
52	TYR_KIT_HUMAN	1pkgA	KIT_HUMAN	279	MG,ADP	BLAminus	-	0
52	TYR_KIT_HUMAN	7khkA	KIT_HUMAN	291	WEJ	BLAminus	-	2
52	TYR_KIT_HUMAN	7khkB	KIT_HUMAN	266	WEJ	BLAminus	-	2
52	TYR_KIT_HUMAN	8pqaC	KIT_HUMAN	299	98A	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqfA	KIT_HUMAN	294	9KI	BLAminus	-	2
52	TYR_KIT_HUMAN	8pq9A	KIT_HUMAN	296	9JI	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqcA	KIT_HUMAN	297	9OO	BLAminus	-	2
52	TYR_KIT_HUMAN	8pqcB	KIT_HUMAN	295	9OO	BLAminus	-	1
52	TYR_KIT_HUMAN	8pq9C	KIT_HUMAN	299	9JI	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqaA	KIT_HUMAN	293	98A	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqdB	KIT_HUMAN	299	9VV	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqgA	KIT_HUMAN	292	9JI	BLAminus	-	5
52	TYR_KIT_HUMAN	8pqdA	KIT_HUMAN	297	9VV	BLAminus	-	0
52	TYR_KIT_HUMAN	8pqgC	KIT_HUMAN	300	9JI	BLAminus	-	0
52	TYR_KIT_HUMAN	1pkgB	KIT_HUMAN	292	MG,ADP	BLAminus	-	13

52	TYR_KIT_HUMAN	8pqbA	KIT_HUMAN	280	9XO	BLAminus	-	11
52	TYR_KIT_HUMAN	8pqeB	KIT_HUMAN	283	9WU	BLAminus	-	13
53	TYR_PDGFRA_HUMAN	8pqiA	PDGFRA_HUMAN	299	9OO	BLAminus	-	10
53	TYR_PDGFRA_HUMAN	8pqhA	PDGFRA_HUMAN	321	9JI	BLAminus	-	0
54	TYR_SYK_HUMAN	5c27A	Peptide	5	50J	BLAminus	-	0
54	TYR_SYK_HUMAN	5c26A	Peptide	6	50H	BLAminus	-	0

See main text Methods for details on the identification of substrate-bound structures in the PDB. This table contains 248 substrate-bound kinase structures identified from the August 2024 PDB.

Supplementary Table 4. Contacts between activation loop residues and substrate

Kinase (Family_Gene_Spec)	PDB	D	F	G	4	5	6	7	8	9		15	14	13	12	11	10	9	8	7	6	5	4	A	P	E
AGC_AKT1	4ekkA				X									X	X	X	X	X	X	X	X	X	X			
AGC_AKT2	1o6kA				X										X	X	X	X	X	X	X	X	X			
AGC_PRKACA	3x2uA	X			X											X	X	X	X	X	X	X	X			
AGC_PRKACA	6mm5E				X										X	X	X	X	X	X	X	X				
AGC_PRKACA	7e12A				X										X	X	X	X	X	X	X	X	X			
AGC_PRKACA	3tnpC				X					X				X	X	X	X	X	X	X	X	X	X			
AGC_PRKACA	6mm7A				X											X	X	X	X	X	X	X	X			
AGC_PRKCI	5li1A	X		X	X									X	X	X	X	X	X	X	X	X	X			
AGC_PRKCI	5lihA			X	X	X									X	X	X	X	X	X	X					
CAMK_CAMK2A	7ujpA				X											X	X	X	X	X	X	X				
CAMK_CAMK2D	2welA				X											X	X	X	X	X	X	X				
CAMK_CAMKII	3kk8A	X			X									X	X	X	X	X	X	X	X	X				
CAMK_CAMKII	5fg8A	X			X								X	X	X	X	X	X	X	X	X	X	X			
CAMK_PHKG1	2phkA			X	X										X	X	X	X	X	X	X	X	X			
CAMK_PIM1	2bzkB				X												X	X	X		X					
CK1_CSNK1D	6ru7A	X	X	X	X	X	X							X		X	X	X	X	X	X					
CK1_CSNK1D	8d7nA	X			X		X							X	X	X	X	X	X	X	X	X				
CMGC_CDK2	1qmzA	X			X		X							X		X	X	X	X	X	X		X			
CMGC_CDK2	3qhrA	X			X									X		X	X	X	X	X	X		X			
CMGC_DYRK1A	2wo6B				X									X		X	X	X	X	X	X	X	X			
OTHER_CDC7	6ya7A	X		X	X		X								X	X	X	X	X		X	X				
OTHER_HASPIN	4oucA				X																					
OTHER_PINK1	8uyhB	X			X		X	X	X	X		X	X	X	X	X	X	X	X	X	X		X			
STE_PAK1	3q4zA	X			X										X	X	X	X	X	X	X	X	X			
STE_PAK4	2q0nA				X									X	X	X	X	X	X	X	X	X	X			
STE_PAK4	4jdhA				X									X	X	X	X	X	X	X	X	X	X			
STE_PAK4	6wlxA				X											X	X	X	X	X	X	X	X			
STE_PAK4	6wlyA				X											X	X	X	X	X	X	X	X			
STE_TNIK	7xzrB	X		X	X											X	X	X	X	X	X	X	X			
TKL_BAK1	3tl8A				X	X	X						X	X	X	X	X	X	X	X	X		X		X	
TKL_IRAK4	8wtfA	X			X							X	X	X	X	X	X	X	X	X	X	X	X			
TKL_IRAK4	4u97A				X							X		X	X	X	X	X	X	X	X	X	X	X	X	
TYR_ABL1	2g2iB															X	X	X	X	X		X				
TYR_CSF1R	3lcdA	X			X								X		X	X	X	X	X	X	X	X	X			
TYR_EGFR	4r3pA	X			X										X	X	X	X	X	X	X	X				
TYR_EGFR	5czhA	X													X	X	X	X	X	X	X	X				
TYR_EGFR	5cziA	X			X										X	X	X	X	X	X	X	X				
TYR_EPHA2	4pdoA															X	X	X	X	X	X	X				
TYR_EPHA3	3fxxA	X			X											X	X	X	X	X	X	X				
TYR_FES	3cd3A													X	X	X	X	X	X	X	X	X				
TYR_FGFR1	3gqiA				X										X	X	X	X	X	X		X				
TYR_FGFR2	2pvfA	X			X											X	X	X	X	X		X				
TYR_FGFR2	3clyA				X										X	X	X	X	X	X	X	X	X			
TYR_FGFR2	5eg3A	X			X											X	X	X	X	X	X	X				
TYR_FGFR3	4k33A	X			X									X	X	X	X	X	X	X		X				
TYR_FGFR3	8uduB	X															X	X	X	X	X	X				
TYR_IGF1R	1k3aA				X							X	X	X	X	X	X	X	X	X		X	X			
TYR_IGF1R	3lvpC				X												X	X	X	X	X	X				
TYR_INSR	1gagA				X							X	X	X	X	X	X	X	X	X	X	X				
TYR_INSR	3bu5A	X			X							X	X	X	X	X	X	X	X	X	X	X	X			
TYR_JAK2-2	6vnbA				X							X	X	X	X	X	X	X	X	X	X	X				
TYR_KIT	1pkgA	X			X										X	X	X	X	X	X	X	X	X			
TYR_PDGFRA	8pqhA	X			X							X	X	X	X	X	X	X	X	X	X	X	X			
TYR_SYK	5c27A	X			X										X	X	X	X	X	X	X	X				

For each kinase, contacts between substrate and activation loops residues (≤ 5 Å) are marked with an "X". A contact with any residue of the DFG motif is listed under "DFG." Residues 4, 5, and 6 of the activation loop are in the adjacent columns. Contacts for the C-terminal region of the activation loop are to the right of the shaded area, starting with residues 15, 14, 13, ..., from the end of the activation loop which typically has the sequence motif "APE" (sometimes "SPE" or "PPE" and other sequences).

Supplementary Table 5. Discrepancies between active-passing PDB structures and AF2 models

Kinase	Explanation for discrepancy with AlphaFold2 model	Potentially False-active structures
CMGC_MAPK8	**Case A:** AF2 model resembles active CMGC_MAPK1 (5v06A).	2no3A, 2h96A, 3o17A, 3o17B, 2no3B, 2g01A, 2h96B, 3o2mA, 3o2mB
AGC_PDPK1	**Case A:** AF2 model resembles the benchmark/phosphorylated structure 4rqkA(<2 Å).	3otuA
CMGC_CDK2	**Case A:** AF2 model resembles the benchmark/substrate-bound structure 1qmzA(<2 Å).	2c4gC, 2wihC, 1vywC, 2bpmC, 2wpaC, 2bkzA, 2bkzC, 3eocA
CMGC_MAPK1	**Case A:** AF2 model resembles the benchmark/phosphorylated structure 5v06A.	6g54A, 4qteA
CK1_CSNK1D	**Case A:** AF2 model resembles the benchmark/substrate-bound structure 6ru6A.	7p7hA
TKL_BMPR1B	**Case B:** The active PDB structure 3MDYC resembles active TKL_TGFBR1 (5friA), but the AF2 model deviates in a critical substrate-binding region of the activation loop.	AF2 model of TKL_BMPR1B (CAMK_DAPK3 template)
CMGC_CLK1	**Case C:** AF2 model resembles the benchmark 2wo6B (<2Å). Active structures >2Å from AF2 model belong to *T. brucei* - likely an orthologous active state.	-
CMGC_DYRK1A	**Case C:** AF2 model resembles the benchmark/substrate-bound 2WO6B. Active structures >2 Å from AF2 model are not likely to interfere with substrate binding, although the 2wo6 form may be preferred.	-
TKL_IRAK4	**Case C:** AF2 model resembles the benchmark 8w3wA (<2Å) and substrate-bound 4u97A. Active structures >2Å from AF2 model are not likely to interfere with substrate binding.	-
TYR_PDGFRA	**Case C:** AF2 model resembles phosphorylated/substrate-bound 3gqiA (TYR_FGFR1). The benchmark/substrate-bound 8pqhA is unphosphorylated, and the AF2 model represents the phosphorylated state.	-
TKL_TGFBR1	**Case C:** Deviation of the AF2 model from the benchmark 5FRIA may not interfere with substrate binding. Several active PDB structures exist with deviations in the same region. This case is less certain than others due to lack of substrate-bound structure to compare with.	-
OTHER_WNK3	**Case C:** AF2 model does not deviate from the benchmark/phosphorylated structure 5o26A at critical substrate-binding sites. Other active structures >2 Å from AF2 model are phosphorylated and resemble the benchmark more closely.	-

Twelve AlphaFold2 models that deviate by more than 2 Å from any of the corresponding active structures in the PDB, and potential explanations for the discrepancies (see text for descriptions of each Case type “A”, “B”, and “C”).

Supplementary Methods. Pseudocode for Kincore-Standalone3 and for code that produces the Kincore database.

1. Read in list of PDB/mmCIF files to be analyzed
2. For each structure:
    a. Extract protein sequence(s) from PDB/mmCIF file for all chains. Sequences are extracted from SEQRES records or _entity_poly.pdbx_seq_one_letter_code records if present. Else they are obtained from the coordinates.
    b. For each chain:
        i. Run hmmsearch of chain sequence against HMMs for each kinase family. HMMs are included for specific kinases that do not align well to the general HMMs (AGC, CAMK, CK1, CMGC, NEK, OTHER, RGC, STE, TKL, TYR), consisting of BUB, EIF2AK41, HASP, MAP3K123, MOS, PAN3, PEAK, PKDCC, PXK, RNASEL, TBCK, TP53RK, ULK, and WNK.
        ii. Determine if the score of the HMM alignment is high enough to be a kinase (>30) or too short <100 residues). Chains which are not kinases are not processed further.
        iii. If the structure is a kinase, identify residues used to determine the conformational state of the structure from the highest scoring HMM. This is performed with the SearchIO function in BioPython. For each HMM, the important residues have been identified by their match-state number in the HMMs. The alignment is used to identify those residues within the kinase sequence/structure. The identified residues consist of: Saltbridge-Lys, Saltbridge-Glu, Glu+4, XHRD residues, XDFG residues, DFG4 and DFG6 residues, APE and APE6-APE12 residues, HPN and HPN7 residues (for the regulatory spine).
        iv. Determine if the kinase chain is a pseudokinase: if the gene name associated with the UniProt identifier (if available) is in a list of 56 human pseudokinase domains (Table S2), then it is labeled "Pseudo." Otherwise, if the HRD-Asp or DFG-Asp are other residue types, the kinase is labeled "Pseudo."
        v. Calculate distances required for the conformational classification, including the Saltbridge NZ/OE distance (for SaltBr-in and SaltBr-out), the LysCB-GluCB distance (for Chelix in or out), the PheCZ-LysCA and PheCZ-Glu4CA (for DFGin, DFGinter, DFGout), the XHRD-DFG6 backbone hydrogen bond distances (save the minimum value, for ActLoopNT-in and out), the APE9CA-ArgO distance (for ActLoopCT-in and out), and the APE10-APE12CB/DFG4-CA distances.
        vi. Calculate dihedral angles for the X-D-F residues for the dihedral angle state determination.
        vii. Determine the spatial state: if PheCZ-Glu4CA≤11 Å and PheCZ-LysCA≥11 Å, then the state is DFGin. If PheCZ-Glu4CA≥11 and PheCZ-LysCA<14, then the state is DFGout. If PheCZ-Glu4≤11 and PheCZ-LysCA<11, then the state is DFGinter. If any atoms are missing or the distances do not fall in the above ranges, the spatial state is labeled 'None'.
        viii. Determine the Chelix and salt bridge states: if the LysCB-GluCB distance is ≤ 10.0 Å, then the Chelix state is Chelix-in. Else it is Chelix-out. If the smaller LysNZ-GluOE distance (between OE1 and OE2) is ≤ 3.6 Å, then the structure is SaltBr-in. Else it is SaltBr-out. If atoms are missing, then the states are labeled 'None'. If the Lys position is not K (as in WNK kinases) or the Glu residue is not E, set the labels to Chelix-na and Saltbr-na (not applicable).
        ix. Determine the dihedral angle states by calculating the dihedral angle distances of the backbone ϕ and ψ angles of the X, D, and F residues between the kinase structure and the centroids of each dihedral angle state, given the spatial state. Each distance has a value of D=2(1-cos(a_input-a_centroid)) where a_input is the input structure dihedral angle and a_centroid is the centroid center dihedral angle. If the structure is in the right spatial state (DFGin, DFGinter, DFGout), the average dihedral angle difference over the six backbone angles is < 0.45 and the chi1 dihedral angle of the Phe side chain is in the right rotamer state (minus, plus, trans), then the structure is assigned to the relevant

state (DFGin-BLAminus, DFGin-BLAplus, DFGin-BLBminus, DFGin-BLBplus, DFGin-BLBtrans, DFGin-ABAminus, DFGout-BBAminus, DFGinter-BABtrans). If any data are missing, the structure is classified as 'None.'

x. Determine the ActLoopNT state. If the XHRD-DFG6 backbone-backbone hydrogen bond distance is ≤3.6 Å, then the structure is ActLoop-in. Else it is ActLoopNT-out. If any atoms are missing, this state is labeled 'None.'

xi. Determine the ActLoopCT state.
   1. Calculate the backbone dihedrals of APE6 and APE7 and test for an intact APE5-helix. If APE6 has values $\phi \in (-180°, 0°)$ and $\psi \in (-100°, 50°)$, and if APE7 has values $\phi \in (-180°, 0°)$ and $\psi \in (-100°, 50°)$, the APE5-helix is considered intact. The APE5-helix is also considered intact if a flip of the APE7-APE6 peptide plane is encountered, indicated by APE6 values $\phi \in (0°, 180°)$ and $\psi \in (-50°, 100°)$ and APE7 values $\phi \in (-180°, 0°)$ and $\psi \in (50°, 180°)$.
   2. Calculate the backbone and χ1 sidechain dihedral angles of the APE8 residue. Active structures are considered with backbone values $\phi \in (-180°, 0°)$, $\psi \in (50°, 180°)$. For non-TYR kinases to be considered active, the conserved Thr[APE8] and Ser[APE8] residues must adopt the *g*− sidechain conformation where $\chi1 \in (-120°, 0°)$ (or 240° to 360° in the shifted coordinate frame computed by adding +360° to the value of χ1 while respecting periodic boundaries at 0° and 360°). If any of these conditions are not satisfied the structure is labeled ActLoopCT-out.
   3. Measure the APE9CA-ArgO distance. If ≤ 6.0 Å for non-TYR kinases and ≤ 8.0 for TYR kinases, then the structure is ActLoopCT-in. Else it is ActLoopCT-out.
   4. For TYR kinases, also check that the APE9 and APE10 backbone dihedrals have values $\phi \in (-180°, 0°)$ and $\psi \in (50°, 180°)$, else the structure is ActLoopCT-out.
   5. If the structure is of a non-TYR kinase, measure the APE10-Cβ/DFG4-Cα distance and check if < 8 Å, else it is ActLoopCT-out.
   6. If the structure is of a non-TYR kinase, measure the APE11-Cβ/DFG4-Cα distance and check if it is in the range (8 Å, 14 Å), else it is ActLoopCT-out.
   7. If the structure is of a non-TYR kinase, measure the APE12-Cβ/DFG4-Cα distance and check if it is in the range (7 Å, 14 Å), else it is ActLoopCT-out.
   8. If the kinase is closest to the PKDCC, TP53RK, or HASP HMMs, then set the label to ActLoopCT-na (not applicable).

xii. Determine the HRD state by calculating backbone dihedrals of the first and second residues of the HRD motif (HRD1 and HRD2). For HRD1 check if $\phi \in (-180°, 0°)$ and $\psi \in (-100°, 50°)$. For HRD2 check if $\phi \in (0°, 180°)$ and $\psi \in (-50°, 100°)$. Else it is HRD-out. If the F-helix Asp is not present in the sequence (DWW motif), then set HRD_label to "HRD-na" (not applicable).

xiii. Determine the activity label of the kinase domain, if the kinase has not been labeled "Pseudo":
   1. Label the domain as Active.
   2. If the spatial_label is not DFGin, then set to Inactive.
   3. If the dihedral_label is not BLAminus, then set to Inactive, unless the assigned HMM is PKDCC, then set to Inactive if the dihedral_label is not ABAminus.
   4. If the HRD_label is HRD-out or HRD-na, set to Inactive.
   5. If neither the Chelix_label or Saltbr_label is "in" or "na", then set to Inactive.
   6. If the current label is Active and the Saltbr_label is "none", set to None.
   7. If the current label is Active or None, and the Saltbr_label is "out", set to Inactive.
   8. If the current label is Active and the ActLoopNT_label is "none", set to None.
   9. If the current label is Active or None, and the ActLoopNT_label is "out", set to Inactive.

10. If the current label is Active and the ActLoopCT_label is “none”, set to None.
11. If the current label is Active or None, and the ActLoopCT_label is “out”, set to Inactive.

xiv. Output the results.