AlphaFold2 models of the active form of all 437 catalytically-competent typical human kinase domains

Bulat Faezov[1,2]

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**

***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). Additionally, there are 57 pseudokinases with the typical kinase domain but without phosphorylation activity. Only 268 of the 437 catalytic typical protein kinases are currently represented in the Protein Data Bank (PDB) in various functional forms. The active form of a kinase must satisfy requirements for binding ATP, magnesium, and substrate. From the structures of 40 unique substrate-bound kinases, as well as many structures with bound ATP, we derived several criteria for the active form of protein kinases. These criteria include: 1) the DFGin position of the DFG-Phe side chain; 2) the "*BLAminus*" conformation based on the backbone and side-chain dihedral angles of the XDFG motif which we previously characterized as required for ATP binding (Modi and Dunbrack, PNAS, 2019); 3) the existence of an N-terminal domain salt bridge between a conserved Glu residue of the C-helix and a conserved Lys of the N-terminal domain beta sheet; 4) backbone-backbone hydrogen bonds of the sixth residue of the activation loop (DFGxxX) and the residue preceding the HRD motif ("X-HRD"); and 5) a contact (or near contact) between the Cα atom of the APE9 residue (9 residues before the C-terminus of the activation loop) and the carbonyl oxygen of the Arg residue of the HRD motif. These last two requirements underscore the structural interplay between the activation loop and the catalytic loop containing the HRD motif that serve to construct a groove capable of binding substrate. With these criteria, only 155 of 437 catalytic kinase domains (35%) are present in the PDB; only 130 kinase domains (30%) are in the PDB with complete coordinates for the activation loop. Because the active form of catalytic kinases is 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. We used active structures we identified from the PDB as templates for AlphaFold2 (AF2) as well as shallow sequence alignments of orthologous kinases from Uniprot (>50% sequence identity to each query) for the multiple sequence alignments required by AF2. We select models for each kinase based on the pLDDT scores of the activation loop residues, demonstrating that the highest scoring models have the lowest or close to the lowest RMSD to 22 non-redundant substrate-bound structures in the PDB. A larger benchmark of 130 active kinase structures with complete activation loops in the PDB shows that 80% of the highest-scoring AlphaFold2 models have RMSD < 1.0 Å and 90% have RMSD < 2.0 Å over the activation loop backbone atoms. We show that several of the benchmark structures from the PDB may be artifacts that are not likely to bind substrate and that the AlphaFold2 models are closer to substrate-bound structures of closely related kinases. Models for all 437 catalytic kinases are available at http://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 (Cohen, Cross et al. 2021). A large majority of human protein kinases take on a common fold first determined by Susan Taylor and colleagues in 1991 (Knighton, Zheng et al. 1991), 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 groove for substrate binding in between the N and C terminal domains. Humans have 481 genes which contain at least one typical full-length protein kinase domain; 13 of these have two kinase domains, for a total of 494 kinase domains (Modi and Dunbrack 2019) [NB: since that paper was published, three kinases have been determined to be pseudogenes]. Of these, 437 are likely catalytic kinases (i.e., participating in phosphorylation of Ser, Thr, or Tyr residues on proteins) and 57 are likely pseudokinases. Currently in the PDB there are structures for 292 typical kinase domains, of which 268 are catalytic kinases and 24 are pseudokinases (Modi and Dunbrack 2022).

Active and inactive conformations of typical kinases have been classified in several ways (Jacobs, Caron et al. 2008, Hari, Merritt et al. 2013, Ung, Rahman et al. 2018, Modi and Dunbrack 2019, Kanev, de Graaf et al. 2021). The active form is generally very similar across kinases because of the requirements of binding ATP, magnesium ions, and substrate. Early on in the history of structure determination of kinases (Levinson, Kuchment et al. 2006), a classification of structures into “DFGin” and “DFGout” was described. 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 under or 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 (Schindler, Bornmann et al. 2000).

There are, however, additional requirements for kinase activity. 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 (Modi and Dunbrack 2019). We found that in addition to being "DFGin," these structures possess specific backbone and side-chain dihedral angles for the DFG motif ("*BLAminus*"), including the backbone dihedral angles of the residue immediately preceding DFG and the side-chain χ_1 dihedral of the DFG-Phe residue. They also possess a well-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 (Yang, Wu et al. 2012). These structures are often referred to as “C-helix-in.” Using these criteria for active kinases, only 183 of 437 catalytic typical human kinase domains are represented in the PDB with active structures. Additional criteria on the positions of the N-terminal and C-terminal segments of the activation loop (see Results) reduce this number to 155 kinases or 35%.

Only 130 of 437 catalytic human kinases (30%) possess active structures and complete coordinates for the activation loop.

The program AlphaFold2 from DeepMind is a deep-learning program for highly accurate protein structure prediction and is trained on a large number of structures from the PDB (Jumper, Evans et al. 2021). It uses as input the query sequence, a multiple sequence alignment (MSA) of homologues of the query, and (optionally), template structures related to the query. DeepMind has provided models of nearly all human proteins produced by AlphaFold2, which are available on a website provided by the European Bioinformatics Institute (Varadi, Anyango et al. 2022). However, only 209 of the 437 (48%) 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, in this paper we describe 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 (Del Alamo, Sala et al. 2022, Heo and Feig 2022). We use similar techniques to compute predicted structures of active kinases.

A key aspect of this work is that we utilize structural bioinformatics to define strict criteria for identifying catalytically active protein kinases structures, including both experimental structures and models predicted by AlphaFold2. We impose criteria on the position of the Phe residue and the dihedral angles of the DFG motif, the formation of the N-terminal domain salt bridge (in kinases that possess the appropriate residues), and on the positions of the N and C terminal halves of the activation loop necessary for the formation of a substrate binding cleft. In addition to reduced MSAs from various sources and active templates from the PDB, we use catalytically active models of kinases produced by AF2 as additional templates for kinases which are more recalcitrant in producing active models and for additional sampling for all kinases. We refer to these as “distillation templates” in analogy with predicted structures that AF2 was trained on (“the distillation training set” (Jumper, Evans et al. 2021)).

We benchmark our protocol with 22 substrate-bound kinase structures in the PDB with complete activation loops and a set of 130 kinase structures that satisfy our active criteria, show that the pLDDT scores for the activation loop are inversely correlated with RMSD of the activation loop for well characterized kinases. With these methods, we produce active models of all 437 catalytic human protein kinase domains. We have made these models available on KinCore (http://dunbrack.fccc.edu/kincore/activemodels).

RESULTS

Catalytic protein kinases

To make active models of all human typical kinases, we need to distinguish between catalytic protein kinase domains and non-catalytic protein kinase domains or pseudokinases. We define *catalytic* protein kinase domains as 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., PAN3, POMK). We previously published an alignment of all 497 human kinase domains from 484 genes annotated by Uniprot (Modi and Dunbrack 2019). This list excludes atypical kinases, such as ADCK, PI3/PI4, Alpha, FAST, and RIO kinases (https://www.uniprot.org/docs/pkinfam.txt). Since that time, three kinase genes have been identified as pseudogenes (SIK1B, PDPK2P, and PRKY) (Frankish, Carbonell-Sala et al. 2023), leaving us with 481 genes and 494 domains.

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 Lysine 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/BUB1R is a pseudokinase (Suijkerbuijk, van Dam et al. 2012); RYK is a pseudokinase (Katso, Russell et al. 1999)). The result of these efforts was a list of 437 active kinase domains in 429 genes. Eight 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 (http://dunbrack.fccc.edu/kincore) and in the text that follows, we use the family name as a prefix in front of each kinase gene followed by the HUGO gene name (e.g., TYR_EGFR) (Seal, Braschi et al. 2023). The catalytic protein kinase domains and associated data are listed in Supplementary Table 1. The pseudokinase domains are listed in Supplementary Table 2.

The characteristics of active protein kinase domains

To identify structural features of the active form of catalytic protein kinases, we created two data sets of structures that constitute likely catalytically active structures of protein kinases. The first consists of structures in the Protein Data Bank of kinases with peptide or protein substrates bound at the active site (**Table 1**). The second consists of structures of 391 catalytic protein kinases with bound ATP or ADP or an ATP analogue that are also in the DFGin "*BLAminus*" conformational state of the DFGmotif that we found characteristic of "catalytically primed" kinase structures (Modi and Dunbrack 2019). The example shown in **Figure 1** is an active form of human AKT1 bound to a substrate, PDB:4ekk (Lin, Lin et al. 2012). To determine what features are important to catalytic activity, we compared the structures in these data

sets to all available structures of kinases in the PDB (without ATP and/or not in the DFGin-*BLAminus* state).

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

Kinase (Family_Gene_Spec)	PDB	Substrate (Unip.)	Auto	Ligand	Len	Sequence	Dihedral label	Salt brdg	DFG6	APE9	Max Spine
AGC_AKT1_HUMAN	4ekkA	GSK3B_HUMAN		ANP	10	GRPRTT**S**FAE	BLAminus	3.0	3.1	4.1	3.9
AGC_AKT2_HUMAN	1o6lA	GSK3B_HUMAN		ANP	10	GRPRTT**S**FAE	BLAminus	2.8	3.1	3.8	3.8
AGC_PRKACA_MOUSE	1l3rE	IPKA_MOUSE		ADP	20	…IASGRTGRRA**S**IHD	BLAminus	2.8	2.9	3.4	4.1
AGC_PRKACA_MOUSE	2qvsE	KAP2_MOUSE		**None**	310	TRRV**S**VCAETF	BLAminus	-	3.0	3.6	4.5
AGC_PRKACA_MOUSE	3idbA	KAP3_RAT		ANP	161	…INRFTRRA**S**VCAEAY…	BLAminus	2.7	3.0	3.4	4.1
AGC_PRKACA_MOUSE	7e0zA	PPLA_MOUSE		ANP	12	TRSAIRRA**S**TIE	BLAminus	3.0	2.9	3.4	4.3
AGC_PRKCI_MOUSE	4dc2A	PARD3_RAT		**ADE**	28	…REGFGRQ**S**MSekrtk…	BLAminus	**5.0**	2.8	3.7	3.9
AGC_PRKCI_HUMAN	5lihA	KPCE_HUMAN		ADP	16	ERMRPFKRQG**S**VRRRV	BLAminus	3.0	3.3	3.6	3.5
CAMK_CAMK2A_HUMAN	7uirA	TIAM1_MOUSE		ATP	19	…HASRMTQLKKQA**A**L	BLAminus	3.5	3.2	3.4	4.0
CAMK_CAMK2D_HUMAN	2welB	KCC2D_HUMAN	A	**K88**	327	MMHRQE**T**VDCLK	BLAminus	**4.7**	3.0	3.6	3.9
CAMK_CAMKII_CAEEL	3kk9B	KCC2D_CAEEL	A	**None**	282	AIHRQD**pT**VDC	BLAminus	**4.9**	2.9	4.1	3.9
CAMK_PHKG1_RABIT	2phkA	Peptide		ATP	7	RQM**S**FRL	BLAminus	2.9	2.9	4.2	4.2
CAMK_PIM1_HUMAN	2bzkB	Peptide		ANP	15	ARKRRRHP**S**GPPTAX	BLAminus	2.7	2.8	3.8	4.0
CK1_CSNK1D_HUMAN	6ru7A	P63_HUMAN		ADP	15	YTPSSASTV**S**VGSSE	BLAminus	2.8	2.9	3.5	4.0
CMGC_CDK2_HUMAN	1qmzA	CDC6_HUMAN		ATP	7	HHA**S**PRK	BLAminus	2.7	3.0	3.2	4.2
CMGC_CDK2_HUMAN	3qhrA	H15_HUMAN		ADP	10	PK**T**PKKAKKL	BLAminus	2.9	2.9	3.3	4.2
CMGC_CLK2_HUMAN	3nr9A	CLK2_HUMAN	A	**NR9**	368	…SRRAK**S**VEDDAE…	BLAminus	2.9	2.8	3.3	3.8
CMGC_DYRK1A_HUMAN	2wo6A	CRUM2_HUMAN		**D15**	8	ARPG**T**PAL	BLAminus	2.7	2.8	3.1	4.2
OTHER_CDC7_HUMAN	6ya7A	MCM2_HUMAN		ADP	15	RRTDALT**X**SPGRDLP	BLAminus	2.8	2.8	3.7	4.0
OTHER_HASPIN_HUMAN	4oucA	H32_HUMAN		**5ID**	12	AR**T**KQTARKSTY	BLAminus	2.6	2.9	-	4.2
STE_PAK1_HUMAN	4zy4A	PAK1_HUMAN	A	**4T3**	329	…PEQSKRS**T**MVGTPYW…	BLAminus	3.2	2.9	3.3	3.9
STE_PAK4_HUMAN	2q0nA	Peptide		**None**	11	RRRRR**S**WYFDG	BLAminus	2.8	2.9	3.6	3.8
TKL_BAK1_ARATH	3tl8A	HPAB2_PSESM		**None**	117	SIDLGE**S**LVQHPL	**ABAminus**	2.8	2.9	3.6	**8.6**
TKL_IRAK4_HUMAN	4u97A	IRAK4_HUMAN	A	**STU**	312	VMTSRIVGT	BLAminus	2.8	-	3.8	3.7
TYR_ABL1_HUMAN	2g2iA	Peptide		ADP	13	AEEEI**F**GEFEAKK	BLAminus	3.3	3.1	7.3	3.7
TYR_CSF1R_HUMAN	3lcdB	CSF1R_HUMAN	A	**BDY**	329	…GNS**Y**TFIDPTQLP…	BLAminus	3.0	3.0	6.9	4.2
TYR_EGFR_HUMAN	5czhA	Peptide		**None**	9	DEED**Y**YEIP	BLAminus	**3.8**	3.0	7.1	4.0
TYR_EPHA2_HUMAN	4pdoA	EPHA2_HUMAN	A	**None**	299	…DPHT**Y**EDPNQAVLK…	**ABAminus**	**6.4**	-	6.5	4.0
TYR_EPHA3_HUMAN	3fxxA	Peptide		ANP	10	kqWDN**Y**EYIW	BLAminus	**4.3**	2.8	6.3	4.0
TYR_FES_HUMAN	3cblA	Peptide		**STU**	6	XI**Y**ESL	BLAminus	2.6	2.7	6.9	3.9
TYR_FGFR1_HUMAN	3gqiA	FGFR1_HUMAN	A	ACP	326	…RPPGLE**F**SFNPSHN…	BLAminus	2.7	3.1	7.1	4.2
TYR_FGFR2_HUMAN	2pvfA	FGFR2_HUMAN		ACP	15	TTNEE**Y**LDLSQPLEQ	BLAminus	2.8	2.9	7.2	4.0
TYR_FGFR2_HUMAN	3clyB	FGFR2_HUMAN	A	ACP	334	…TTNEE**Y**LDLsqpled…	BLAminus	2.6	2.9	6.9	4.1
TYR_FGFR3_HUMAN	4k33B	FGFR3_HUMAN	A	ACP	325	…PPGLD**Y**SFDTSkppe…	BLAminus	2.9	3.1	6.9	4.0
TYR_IGF1R_HUMAN	1k3aA	IRS1_HUMAN		ACP	14	KKKSPGE**Y**VNIEFG	**ABAminus**	**4.8**	2.9	7.1	3.9
TYR_IGF1R_HUMAN	3lvpB	IGF1R_HUMAN		**PDR**	336	…YETDY**Y**RKGGKGLLP…	**ABAminus**	**5.7**	3.1	7.1	4.0
TYR_INSR_HUMAN	1ir3A	IRS1_HUMAN	A	ANP	18	…PATGD**Y**MNMSPVGD	BLAminus	**4.4**	2.9	6.8	3.8
TYR_INSR_HUMAN	3bu5A	IRS2_MOUSE		ATP	15	AYNPYPED**Y**GDIEIG	BLAminus	3.1	2.8	7.4	4.0
TYR_KIT_HUMAN	1pkgB	KIT_HUMAN		**None**	320	…NN**X**VXIDPTQLP…	BLAminus	2.9	2.9	6.5	3.9
TYR_SYK_HUMAN	5c27A	Peptide		**50J**	5	EV**Y**ES	BLAminus	2.8	3.0	6.8	3.8

Autophosphorylation complexes are marked with "A" in column 4. The phosphorylation site is in bold red type. 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. DFG6 is the shorter backbone-backbone hydrogen bond distance of the sixth residue of the activation loop (DFGxxX) and the residue before the HRD motif (Xhrd). APE9 is the distance between the Cα atom of the 9th residue from the end of the activation loop (XxxxxxAPE) and the backbone carbonyl oxygen of the Arg residue of the HRD motif. The Max Spine distance is the largest of the three spine distances of the regulatory spine of Kornev and Taylor (Kornev and Taylor 2010). Each spine distance is the closest atom-atom distance of any pair of side chain atoms in two neighboring spine residues.

Table 1 presents a list of unique kinase-substrate and kinase-pseudosubstrate complexes in the PDB and some structural parameters that will be considered below. Some of the “substrates” are in fact substrate-mimicking inhibitors, which bind very similarly to substrates. Some kinases are represented more than once if they contain different bound substrates in the active site. Eleven of the 40 complexes are “autophosphorylation complexes,” which we previously identified as homodimeric complexes in crystals of protein kinases in which a known autophosphorylation site of one monomer sits in the active-site and substrate-binding groove of another monomer in the crystal (Xu, Malecka et al. 2015). These include autophosphorylation complexes of sites in the activation loop (STE_PAK1, 4zy4; TYR_IGF1R, 3lvp) and the kinase insert loop (TYR_FGFR1, 3gqi; TYR_FGFR3, 4k33). The remainder are N or C

terminal tails (CAMK_CAMK2D, 2wel; CAMK_CAMKII, 3kk9; CMGC_CLK2, 3nr9, TYR_CSF1R, 3lcd; TYR_KIT, 1pkg; TYR_EPHA2, 4pdo; TYR_FGFR2, 3cly). Three other complexes are with larger proteins which are either direct substrates or inhibitors or both (AGC_PRKACA:KAP2, 2qvs; AGC_PRKACA:KAP3, 3idb; TKL_BAK1:HPAB2, 3tl8). The last of these is a plant kinase/pathogen-inhibitor complex (Cheng, Munkvold et al. 2011). The autophosphorylation complexes (marked with "A" in column 4 of Table 3) and inhibitor protein complexes provide insights of how kinases phosphorylate amino acids in the context of folded protein domains, as opposed to intrinsically disordered regions (IDRs).

We previously identified several criteria for active structures in the PDB for catalytic protein kinase domains (Modi and Dunbrack 2019): 1) the spatial label must be *DFGin*; 2) 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°); 3) 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 this paper, we validate these criteria and extend them 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, which must be positioned for binding a substrate, as determined by a residue 9 positions from the end of the activation loop. We also consider the presence of the regulatory spine defined by Kornev and Taylor. We review each of these in turn.

DFGin conformation

The position of the DFG-Phe residue determines, in part, the position of the catalytic DFG-Asp residue. We defined DFGin by the distance between the DFG Phe Cζ atom and the Cα atoms of two residues in the N-terminal domain (Modi and Dunbrack 2019): 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 ATP and non-ATP-bound structures in **Figure 2**. The vast majority of ATP-bound structures (defined as having ligands with PDB 3-letter codes: ATP, ADP, ANP, or ACP in the active site) are *DFGin* with LysCα-PheCζ distance > 11 Å and Glu4Cα-PheCζ distance < 11 Å. All of the substrate-bound structures listed in Table 1 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 (usually HRD, but Tyr in AKT1). This hydrogen bond helps position the catalytic aspartic acid residue of the YRD motif, which interacts with the Ser or Thr hydroxyl atoms of substrate residues to be phosphorylated. The *BLAminus* conformation is required for these interactions (Modi and Dunbrack 2019). *ABAminus* structures involve a "peptide flip" of the X-D residues, such that the carbonyl of the X residue points upwards and does not interact with Y/H of the Y/HRD motif. Many of these structures are missolved and should be *BLAminus*, as demonstrated by poor electron density for the X residue carbonyl oxygen (Modi and Dunbrack 2019). The backbone and side-chain dihedral angles of the XDFG motif for the substrate-bound structures in Table 1 are shown in Figure 3.

In addition, 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. The minus rotamer of the Phe side chain is required for this interaction, since the plus rotamer of the inactive BLAplus and BLBplus conformations points upwards (instead of downwards in *BLAminus*) and pushes the C-helix outwards, breaking the salt bridge (Modi and Dunbrack 2019).

While 69% of ATP-bound and 65% of non-ATP-bound catalytic kinase structures are in the *BLAminus* conformation, the role of the *BLAminus* configuration becomes clearer when combining it with the formation of the N-terminal domain salt bridge. In **Figure 4A**, the density of distances of the salt bridge atom pairs (Nζ in the β3 Lys residue, Oε_1 or Oε_2 in the C-helix Glu residue (whichever is shorter)) is plotted for *BLAminus* and non-*BLAminus* structures with and without ATP. When *BLAminus* structures are bound with ATP, the salt bridge is strongly favored with a mean distance of about 3.0 Å (upper left of Figure 4A). However, ATP-bound structures that are not in the *BLAminus* state have a broken salt bridge, with most structures having a Lys/Glu distance greater than 10 Å (lower left panel of Figure 4A). Even in the absence of ATP, the *BLAminus* conformation encourages the formation of the salt bridge (upper right vs lower right panels of Figure 4A).

Conversely, if we require salt bridge formation ("SaltBr-In") with a cutoff of Nζ/Oε distance of 3.6 Å, 99% of ATP-bound structures are in the *BLAminus* conformation. When the salt bridge is not formed ("SaltBr-Out"), only 19% of the structures are *BLAminus* (**Figure 4B**).

ActLoopNT

We examined the substrate-bound structures listed in Table 1 for further 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 must be in the appropriate position for forming a substrate binding groove. Examples from four families are shown in **Figure 5** with the activation loops in magenta, phosphorylated residues in the activation loop in pink, ATP (or analogs) in green sticks, and the substrates in blue.

Table 2. Contacts between activation loop residues and substrate

Kinase (Family_Gene_Spec.)	PDB	Len	DFG	4	5	6		15	14	13	12	11	10	9	8	7	6	5	4	APE
AGC_AKT1_HUMAN	4ekkA	10		X						X	X	X	X	X	X	X	X	X	X	
AGC_AKT2_HUMAN	1o61A	10		X							X	X	X	X	X	X	X	X	X	
AGC_PRKACA_MOUSE	1l3rE	20	X	X							X	X	X	X	X	X	X	X	X	
AGC_PRKACA_MOUSE	2qvsE	310		X						X	X	X	X	X	X	X	X	X	X	
AGC_PRKACA_MOUSE	3idbA	161		X						X	X	X	X	X	X	X	X	X	X	
AGC_PRKACA_MOUSE	7e0zA	12		X								X	X	X	X	X	X	X	X	
AGC_PRKCI_MOUSE	4dc2A	28		X							X	X	X	X	X	X	X	X	X	
AGC_PRKCI_HUMAN	5lihA	16	X	X	X						X	X	X	X	X	X	X			
CAMK_CAMK2A_HUMAN	7uirA	19	X	X								X	X	X	X	X	X	X		
CAMK_CAMK2D_HUMAN	2welB	327		X								X	X	X	X	X	X	X		
CAMK_CAMKII_CAEEL	3kk9B	282	X	X								X	X	X	X	X	X	X		
CAMK_PHKG1_RABIT	2phkA	7	X	X							X	X	X	X	X	X	X	X		
CAMK_PIM1_HUMAN	2bzkB	15		X								X	X	X	X		X			
CK1_CSNK1D_HUMAN	6ru7A	15	X	X	X	X				X		X	X	X	X	X				
CMGC_CDK2_HUMAN	1qmzA	7	X	X		X				X		X	X	X	X	X	X		X	
CMGC_CDK2_HUMAN	3qhrA	10	X	X						X		X	X	X	X	X	X		X	
CMGC_CLK2_HUMAN	3nr9A	368										X	X	X	X	X	X	X	X	
CMGC_DYRK1A_HUMAN	2wo6A	8		X						X		X	X	X	X	X	X	X	X	
OTHER_CDC7_HUMAN	6ya7A	15	X	X		X				X		X	X	X	X	X		X	X	
OTHER_HASPIN_HUMAN	4oucA	12		X				-	-	-	-	-	-	-	-	-	-	-	-	-
STE_PAK1_HUMAN	4zy4A	329	X	X										X	X		X	X		
STE_PAK4_HUMAN	2q0nA	11		X						X	X	X	X	X	X	X	X	X	X	
TKL_BAK1_ARATH	3t18A	117		X								X	X	X	X	X				
TKL_IRAK4_HUMAN	4u97A			X						X	X	X	X	X	X	X	X		X	
TYR_ABL1_HUMAN	2g2iA	13								X		X	X	X	X	X		X		
TYR_CSF1R_HUMAN	3lcdB	329	X	X							X	X	X	X	X	X	X	X	X	
TYR_EGFR_HUMAN	5czhA	9	X								X	X	X	X	X	X	X	X		
TYR_EPHA2_HUMAN	4pdoA	299										X	X	X	X	X	X	X		
TYR_EPHA3_HUMAN	3fxxA	10		X									X	X	X	X	X	X		
TYR_FES_HUMAN	3cblA	6		X						X	X	X	X	X	X	X		X		
TYR_FGFR1_HUMAN	3gqiA	326		X							X	X	X	X	X	X		X		
TYR_FGFR2_HUMAN	2pvfA	15	X	X									X	X	X	X		X		
TYR_FGFR2_HUMAN	3clyB	334				X				X	X	X	X	X	X	X	X	X		
TYR_FGFR3_HUMAN	4k33B	325	X	X						X	X	X	X	X	X	X		X		
TYR_IGF1R_HUMAN	1k3aA	14		X				X	X	X	X	X	X	X	X	X		X	X	
TYR_IGF1R_HUMAN	3lvpB	336		X	X	X							X	X	X	X		X		
TYR_INSR_HUMAN	1ir3A	18	X	X						X	X	X	X	X	X	X	X			
TYR_INSR_HUMAN	3bu5A	15	X	X				X	X	X	X	X	X	X	X	X	X	X	X	
TYR_KIT_HUMAN	1pkgB	320	X	X								X	X	X	X		X	X		
TYR_SYK_HUMAN	5c27A	5	X	X							X	X	X	X	X	X	X			

For each kinase, contacts between substrate and activation loops residues 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" (often "SPE" or "PPE").

Besides the conformation of the DFG motif (the Phe/Tyr residue of DFG is shown in orange sticks), 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 (yellow in each figure), have similar conformations and positions across the members of each family. The second is that in the Ser/Thr kinases of the AGC, CAMK, and CMGC families, the C-terminal segment of the activation loop also shares a common conformation and position across family members. In Figure 5, residues 8 and 9 from the end of the activation loop are shown in cyan. The conformation of residues 8-11 from the end of the

activation loop resemble the hull shape of an upside-down, round-bottom boat. Residues 8-9 in Tyr kinases are also in a common position, although the structure diverges in residues 10 and 11 more than in the Ser/Thr kinase members. In Tyr kinases, the substrate binds directly to these residues in the form of a short beta sheet (blue lines in Figure 5, lower right). The conformation may diverge to accommodate substrates with larger or smaller side chains.

We determined which residues within the activation loop form direct contacts with substrate residues (any atom contact within 5 Å between substrate residues and the DFG…APE sequence). The results are shown in **Table 2**. Most substrates have a contact with one or more of the DFG residues as well as the fourth residue of the activation loop, while a small number have contacts with residues 5 and 6. By looking at the structures, we identified the existence of backbone-backbone hydrogen bonds between residue 6 of the activation loop (DFGxx***X***) and the residue immediately preceding the HRD motif (***X***HRD) (**Figure 6A**). We used this distance previously to characterize active structures in the PDB (Modi and Dunbrack 2019). This hydrogen bond is present in all of the substrate-bound structures in Table 1 ("DFG6" in the table) except for IRAK4 (PDB:4u97) where the DFG6 residue is disordered. Almost 99% of *BLAminus* structures (386 of 391) with bound ATP contain this hydrogen bond with the minimum N-O or O-N backbone-backbone distance less than 3.6 Å (**Figure 6B**, upper left panel). The only exception is an activation-loop swapped structure of STE_MAP4K1 (PDB: 6cqd) in which the N-terminal portion of the activation loop forms an α-helix. Even without ATP, 97% of *BLAminus* structures contain the DFG6/XHRD hydrogen bond (Figure 6B, upper right). In the non-*BLAminus* state, only 24% of structures contain this hydrogen bond (Figure 6B, bottom panels).

ActLoopCT

The conformation of the C-terminal end of the activation loop is critical for binding substrate. Most substrate-bound structures in Table 2 contain contacts between the substrate and residues 4-11 from the end of the activation loop, which ends in the sequence motif “APE.” 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. This contact is shown in 23 non-TYR kinase structures from Table 2 in **Figure 7A**. The Cα-O distance is ≤ 4.2 Å in all of these structures.

Aurora A kinase (AURKA) is a good example of the utility of these contacts. In the *BLAminus* state, there are two dominant conformations of the entire activation loop of AURKA. **Figure 7B** (left panel) shows five structures that contain these contacts. This comprises seven 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. Most *BLAminus* structures of AURKA, however, resemble the structures

shown in Figure 7B (right panel). 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 AGC and CAMK families. In the active structures, the Cα-O distances are about 3.6 Å, while in the inactive structures, the distance is more than 10 Å.

In Table 1, the APE9(Cα)-hRd(O) distance ranges from 3.4 to 4.2 Å in the substrate complexes in the Ser/Thr kinases (all families except TYR). This suggests that the Cα-O interaction is a CH-O hydrogen bond, which have been observed in proteins (Derewenda, Lee et al. 1995). In 271 of 355 non-TYR catalytic kinases, the APE9 residue is a glycine, which forms Cα-O hydrogen bonds more readily than other amino acids for steric reasons. In the TYR family kinases in Table 2, the APE9(Cα)-hRd(O) distance is longer and ranges from 6.5 to 7.4 Å.

We examined the distributions of this distance in ATP-bound and non-ATP structures in the *BLAminus* and other conformational states (**Figure 8**). For non-TYR kinases, the APE9-hRd distance is typically less than 6 Å in *BLAminus*/ATP-bound structures (Figure 8A, upper left panel), while the distance is much greater than 6 Å in a majority of non-*BLAminus* structures (Figure 8A, lower panels). Many of the *BLAminus*/ATP-bound structures with longer APE9-hRd distances are AURKA structures, such as those in Figure 7B. As with the substrate bound structures, this distance is somewhat longer for *BLAminus*/ATP-bound structures of TYR kinases than for non-TYR kinases, ranging from 5 to 8 Å (Figure 8B, upper left panel), The large peak at 5 Å are all structures of FGFR2. For other kinases, the distance is typically between 6 and 8 Å. One third of non-*BLAminus* TYR kinase structures have an APE9-hRd distance greater than 8 Å.

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-termina 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 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 1**, 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 affect 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 2**. 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). 96% of BLAminus structures contain an intact spine. Most of those with a broken spine occur because of the position of the Glu4 or HPN7 residues, neither of which interact with the substrate or ATP. In general, the Spine distance does not add to the other criteria described above, so for the sake of simplicity, we do not use it 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 at those capable of binding ATP, Mg ions, and substrate, with the following criteria:

1. DFGin spatial state
2. *BLAminus* dihedral angle state
3. SaltBr-In state (Nζ/Oε distance < 3.6 Å)
4. ActLoopNT-In (DFG6-Xhrd backbone hydrogen bond < 3.6 Å)
5. ActLoopCT-In (APE9-Cα/hRd-O distance < 6 Å in non-TYR kinases and < 8 Å in TYR kinases)

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 AlphaFold2 models with all of Uniprot90 as the MSA sequence database reproduce this unusual feature even in *BLAminus* structures. 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 (Modi and Dunbrack 2019), 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.

We updated Kincore-standalone to calculate the relevant data for all human kinases. The results for the PDB are shown in Table 3 for catalytic kinases. Of 437 real kinase domains in the human proteome, only 155 (35.5%) have active structures in the PDB. Of these, only 130 have complete sets of coordinates for the backbone of the activation loop, comprising less than 30% of catalytic kinases in the human proteome. We therefore chose to see if we could use AlphaFold2 to produce active structures of all 437 real kinase domains in the human proteome.

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

	With/without ActLoop disorder			With no ActLoop disorder		
	Chains	Catalytic human kinases	Percent (of 437)	Chains	Catalytic human kinases	Percent (of 437)
Any conformational state	8277	268	61.3	4640	217	49.7
DFGin	7097	252	57.7	4221	196	44.9
DFGin+BLAminus	4489	202	46.2	3261	160	36.6
DFGin+BLAminus+SaltBr-in	3644	188	43.0	2768	147	33.6
DFGin+BLAminus+actloopNT-in	4319	193	44.1	3201	158	36.2
DFGin+BLAminus+actloopCT-in	3675	162	37.0	2934	141	32.3
Active	**3013**	**155**	**35.5**	**2531**	**130**	**29.7**

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 paralogues) for each kinase were created from UniProt such that each sequence in an orthologue 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 orthologue set was filtered with CD-HIT so that no two sequences in the set were more than 90% identical to each other. This was done to create diversity within the orthologue sets for each kinase. We also created “Family” sequence sets consisting of all the human kinase domains within each kinase family.

To create a template set, we identified all active structures of catalytic kinases in the PDB (including non-human kinases) using the criteria given above and selected two structures from different PDB entries (if available) with the largest number of coordinates for the activation loop residues (to select in favor of complete activation loops). If more than two structures were available with the same number of ordered residues in the activation loop, those with the highest resolution were selected. This resulted in a set we named "ActivePDB," consisting of 165 kinase domains from 278 PDB entries.

We applied AlphaFold2 to all 437 human catalytic kinase domains, using the orthologue and family sequence sets and the ActivePDB template set. Different depths of the sequence alignment were utilized ranging from 1 sequence to 90 sequences. Only two of AlphaFold2's five models utilize templates, so only models 1 and 2 were run when templates were included in the calculations. The models were relaxed with AMBER and the standard AlphaFold2 protocol, and we assessed the activity state of both the unrelaxed and relaxed models. In many cases, hydrogen bonds that were broken in the unrelaxed models were formed properly in the relaxed models.

We downloaded structures of all 437 kinases from the EBI website of AlphaFold2 models. Only 208 of the 437 kinase domains contain active structures within this set. When we ran all five models within AlphaFold2 with default parameters (with and without templates, Uniprot90 as the sequence database), we obtained active models of 281 catalytic kinases (out of 437) using the PDB70 template database and active models of 298 catalytic kinases using no templates. By comparison, under different conditions, using the ActivePDB templates and Distillation templates, orthologue and family sequence databases, and different MSA depths, we obtained between 371 and 421 active kinases (**Figure 9**), depending on the input template and MSA data sources.

No one set of inputs (MSA source, MSA depth, template database) produces active models of all 437 catalytic targets but combining the models from the different sets achieved models of 435 out of 437 targets. For two kinases, we needed special procedures. 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. It is possible that OBSCN-2 is a pseudokinase.

For LMTK2, all AlphaFold2 models made from the runs in Figure 9 formed a folded activation loop containing a strand-turn-strand motif that would be inconsistent with substrate binding. This structure forms in many DFGout structures of TYR kinase family members. We added additional distillation templates to the distilledAF2 template set of active structures of TYR_AATK (also known as LMTK1) and TYR_LMTK3. This produced active models of LMTK2 with very shallow sequence alignments (1-3 sequences from the orthologue data set). LMTK2 has been shown to phosphorylate CFTR and other substrates involved in neuronal activity (Luz, Cihil et al. 2014).

In Table 4, we show the number of active kinase domains produced by different combinations of template database and MSA source summed over the MSA depths run for each combination shown in Figure 9. Using all the models with Uniref90 sequences produced active models of only 308 kinase domains. The ActivePDB template set plus the Family models and Ortholog models combined produced active models of 435 kinases. The only two kinases that required the distillation templates were LMTK2 and LMTK3; they only formed active models with 5 or fewer sequences from the ortholog set. As noted above, LMTK2 required the LMTK3 model, effectively a redistillation template. However, the quality of models in terms of pLDDTs of the activation loop is improved by including the distillation set models, as we show in the next section.

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
PDB70 (EBI)	Uniref90 (EBI)	209
PDB70	Uniref90	281
None	Uniref90	298
All (PDB70 and None)	Uniref90	308
ActiveAF2	Family	426
ActiveAF2	Ortholog	429
ActiveAF2	All	431
ActivePDB	Family	431
ActivePDB	Ortholog	431
All	Family	432
All	Ortholog	435
ActivePDB	All	435
All	All	437

The first line of the table is derived from data from the EBI database of Alphafold2 structures.

The structures of substrate-bound kinases (Figure 5) show that in active kinases, the activation loop is generally situated against the kinase domain, extending from the DFG motif towards the right-edge of the kinase domain (as generally pictured in Figure 5). 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 AlphaFold2 models produced by our pipeline. Dozens of examples of active AF2 models are shown in **Figure 10** for the AGC, CMGC, STE, and TYR kinase families.

Picking the best model with pLDDTs scores of the activation loop

To benchmark the behavior of our pipeline in modeling active structures of catalytic kinases, we first compared the collection of AlphaFold2 models for the 22 kinases listed in Table 1 that have complete activation loops with their experimental 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).

These experimental structures contain substrates so are likely to be one (of possibly several) 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 the pLDDT values of the activation loop to pick out good models that pass our "Active" tests described above. We also wanted to know if the distillation models provided better models of active structures in some cases.

In **Figure 11**, we show scatterplots of RMSD vs pLDDT of the activation loop for these 22 kinases. The results demonstrate that for most of the kinases, the highest pLDDT for the activation loop (defined as the minimum pLDDT value over the activation loop residues in the model) also produced the best or very close to the best RMSD to the structures listed in Table 1. The distillation templates ("ActiveAF2") produced significantly better models than the ActivePDB templates for CMGC_CDK2 and CAMK_PIM1, and higher pLDDTs for most kinases. Thus, it seems likely that the extra sampling with the distillation templates may produce better models or more confident models for active structures of kinases.

To extend the benchmark, we picked out at least one structure for each of the 130 human catalytic kinases with active structures in the PDB and complete activation loops. When all or almost all of the structures for a particular kinase were similar (except for perhaps a few outliers), we picked out only one structure as a representative. When more than one conformation was represented in multiple PDB entries, we picked out a representative from each, labeling them "conf1," "conf2,", etc. The structures labeled "conf1" were generally those that most closely resembled the substrate-bound structures in Table 1. The distribution of RMSD for the highest scoring models (highest min(pLDDT over the activation loop)) and the distribution of min_pLDDT values for the conf1 structures are shown in **Figure 12**. The results show that 104 (80%) of the 130 kinases are represented by a model with less than 1 Å backbone atom RMSD (N,CA,C,O) over the whole activation loop (after superposition of the C-terminal domains of each kinase). A total of 117 (90%) are less than 2.0 Å.

We can show that when multiple conformations of the activation loop of a given kinase are considered "Active", our AlphaFold2 models are generally close to one of them, and this structure most closely resembles the substrate-bound structures in Table 1. By visually clustering the structures of human CDK2 that pass our criteria, we identified four predominant conformations (**Figure 13A**). The

conf1 benchmark structure (PDB: 1QMZA) is a substrate-bound structure listed in Table 1. Structures very similar to the conf1 structure are also the only ones that are phosphorylated on residue T160. For conformations conf2 (2BZKA), conf3 (5UQ1A), and conf4 (1FINA), the closest structures among the AlphaFold2 models have RMSD of 2.59 Å, 1.09 Å, and 2.78 Å respectively, all with min_pLDDT of less than 50.0. This contrasts with the best model of conf1, which has an RMSD of 0.48 Å to PDB:1QMZA and min_pLDDT of 84.2.

SRC presents an interesting example (**Figure 13B**). The human SRC structures which are "Active" by our criteria and contain fully ordered activation loops (PDB: 1Y57A (green in Fig 16B), 1Y16A, 1Y16B (orange in Fig 13B)) do not resemble substrate-bound structures of Tyr kinases in Table 1, such as ABL1 (PDB:2G2I). However, there is a structure of chicken SRC in the PDB (PDB 3DQW, chains A-C, blue in Figure 13B) that is quite similar to PDB 2G2I listed in Table 1 (Figure 13B, magenta). 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. Thus, there is no substrate-binding-capable structure of the SRC kinase domain annotated as human in the PDB, while the chicken SRC structure (PDB: 3DQW, chain A) presents a suitable model of active SRC. This is an important observation because an inactive structure of human SRC that is incapable of binding substrate (PDB: 1Y57) is often used as the basis of molecular dynamics simulations of the *active* protein (Foda, Shan et al. 2015, Fajer, Meng et al. 2017, Joshi, Burton et al. 2020). In the benchmark of 130 active structures, we replaced the structure of human SRC (PDB: 1Y57A) with that of chicken SRC (PDB: 3DQWA).

MAPK1 has two main conformations in our benchmark (Figure 13C). One of them resembles the substrate-binding structures in Table 1 (Figure 13C, blue, left panel). The other has a bulge in the activation loop in the C-terminal half that places the activation loop over the APE motif and in contact with the G-helix (Figure 13C, orange, left panel). AlphaFold2 reproduces both of these structures almost exactly (Figure 13C, right panel) with RMSD of ~0.3 Å in both cases. The active models are produced by the ActivePDB templates, while the alternate conformation models are produced by the ActiveAF2 distillation templates. The distillation templates included a structure of the closely related CMGC_MAPK3, which also has the same bulge as the orange structures in Figure 13C. The benchmark structure of MAPK3 (PDB: 4QTBA) in fact has the same bulge. However, the highest scoring AlphaFold2 model resembles substrate bound structures with an RMSD of 5.1 Å and is one of the RMSD outliers in Figure 12A.

In addition to CMGC_MAPK1 and CMGC_MAPK3 just discussed, there are 11 other kinases where the highest pLDDT models have more than 2 Å RMSD to the structural representative we chose These kinases are: CMGC_HIPK3 (2.01 Å RMSD), CMGC_MAPK7 (5.74 Å), OTHER_BUB1 (3.49 Å), OTHER_WNK3 (2.72 Å), STE_MAP3K14 (3.97 Å), STE_MAP3K5 (2.42 Å), STE_STK3 (2.30 Å), STE_TNIK (2.47 Å), TKL_ACVR1 (2.49 Å), TKL_ACVR2B (2.39 Å), and TKL_BRAF (2.53 Å). These are analyzed and discussed in **Supplementary Figure 3** and **Supplementary Figure 4**. In some cases, the

model and PDB structure only differed in the outermost residues of the activation loop (from the DFG and APE motifs). This occurred for STE_STK3, STE_TNIK, TKL_ACVR2B, and OTHER_WNK3 for example. In TKL_ACVR2B, there is a change in position of residues 10-17 or a 30-residue activation loop, while residues 1-9 and 18-30 are very similar in the benchmark structure (2QLUA) and the AF2 models.

In some other cases, the AF2 structure appears capable of binding substrate while the PDB structure does not. In some cases, this can be demonstrated by comparing the benchmark structure to that of closely related kinases in the PDB and in our AF2 models. For example, the CMGC_HIPK3 and CMGC_HIPK2 benchmark structures are quite different in the C-terminal region of the activation loop. The AF2 models of HIPK2 and HIPK3 closely resemble the HIPK2 conformation (PDB:7NCFA) but not the HIPK3 experimental structure (PDB: 7O7IA). Similarly, for TKL_ACVR1A, the PDB structure (6UNSA) blocks the active site, while the AF2 models resemble the TKL kinase BAK1 (PDB:3TL8A) from Table 1, which contains a substrate peptide.

For some other kinases, the AF2 models have poor pLDDT of 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. For all three kinases in the RAF family (ARAF, BRAF, RAF1 or CRAF), the min_pLDDT scores for the activation loop are below 40. For BRAF, the top scoring AF2 models are not very similar to the benchmark structure with an RMSD of 2.53 Å (PDB:4MNEB, the only structure of BRAF with a complete activation loop that passes our "Active" criteria). It is not known if this PDB structure is fully capable of binding substrates or whether the AF2 models are in fact better models of substrate-capable structures.

DISCUSSION

We have developed a structural bioinformatics approach to identifying structures of typical protein kinases that are likely capable of binding ATP, metal ions, and substrates and catalyzing protein phosphorylation, which is involved in nearly all cellular processes in eukaryotes. We applied these criteria to experimental structures, which enabled us to develop a set of templates that could be used to model all 437 catalytic protein kinases in their active form with AlphaFold2. The same criteria enabled us to distinguish active structures among the models produced by AlphaFold2, which we cycled back into the protocol as templates for producing additional models with improved pLDDT scores. We refer to these as distillation templates, in analogy to the distillation models that the team at DeepMind used as additional training data for the original implementation of AlphaFold2. We demonstrate that the models with the highest values of pLDDT for the activation loop residues also most closely resemble substrate-bound structures of kinases in the PDB.

While much attention has been given to the structure of the active site residues surrounding ATP, including the DFG motif and the N-terminal domain salt bridge, we examined substrate-bound structures of protein kinases in the PDB to define criteria that ensure the presence of a substrate binding site

necessary for the phosphorylation reaction. In substrate-bound structures, the activation loop is extended away from the ATP binding site, lying against the surface of the kinase domain. To accomplish this, the activation loop interacts with the relatively fixed positions of residues in the catalytic loop in and around the HRD motif. This occurs both near the N-terminus of the activation loop in a backbone-backbone hydrogen bond of residue 6 of the activation loop with the residue that immediately precedes the HRD motif, and near the C-terminus of the activation loop where the Cα atom of residue 9 from the end of the loop makes a short contact with the backbone carbonyl of the Arg residue of the HRD motif. While other distances could also be used as criteria, we found that all substrate-bound structures in the PDB satisfy these two rules and that the vast majority of experimental and computed structures that satisfy these criteria appear to form a functional substrate binding site. For some kinases, there remains some conformational diversity of the activation loop after satisfying these criteria. In some cases, multiple conformations of the outer portion of the activation loop may be capable of phosphorylating substrates.

In other cases, some conformations that satisfy our criteria may block substrate binding. Unfortunately, there does not seem to be a readily identifiable criterion that would be applicable across kinases to identify such situations. This phenomenon does seem to be rare. For example, MAPK1, MAPK3, MAPK7 share an alternate conformation in experimental structures that would block substrate binding. AlphaFold2 produces these structures but also substrate-capable structures that resemble substrate-bound structures in the same CMGC family. These models are the ones we have made available in a set of models of active structures of all 437 catalytic typical protein kinases in the human proteome (http://dunbrack.fccc.edu/kincore/activemodels).

We have found that AlphaFold-Multimer is in some cases capable of making models of substrate-bound structures of typical protein kinases when given a peptide substrate and Uniref90 as a sequence database. But it is not always able to do make an active model of the kinase activation loop without appropriate templates and shallow sequence alignments. But doing this sometimes disrupts its ability to place the substrate in the active site, probably due to the lack of sequence information for the substrate MSA. 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

Orthologue sequence sets

We first searched UniProt for Pfams PF00069 and PF07714 to collate 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 (Modi and Dunbrack 2019), we used PSI-BLAST 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 (Fu, Niu et al. 2012) to create lists of orthologues (or close paralogues) with no more than 90% sequence identity to each other. These sequences were used as query databases in AlphaFold2 calculations.

AlphaFold2

We used DeepMind's advanced machine learning model, AlphaFold2, to predict the structures of proteins in the kinase family and their orthologs. The code for AlphaFold2 was sourced from DeepMind's official GitHub repository (https://github.com/deepmind/alphafold). The computations were performed on workstations with NVIDIA GeForce GPUs (8, 12, or 24 Gbytes each). Each system was equipped with Linux (Ubuntu 20.04), CUDA11, Python 3.8, and TensorFlow 2.3.1.

Data Input and Preparation: Three sets of sequence databases were used to create multiple sequence alignments: the default UniRef90 database, an additional kinase family-focused sequence database (all 496 human kinases in the human proteome, separated into each family), and a kinase orthologs-focused sequence database (described above). Templates for the analysis were sourced from the default PDB70 set, a curated selection of active PDB models identified through our criteria by Kincore ("ActivePDB"), and a distilled set of AlphaFold2 models that passed Kincore criteria with activation loop pLDDT scores of 60 or higher ("ActiveAF2" or "distilled").

Model Configuration and Implementation: Calculations with AlphaFold2 were conducted using the recommended configurations provided by DeepMind. The multiple sequence alignment was prepared using the hh-suite package (Steinegger, Meier et al. 2019) and subsequently fed into the model for structure prediction. When using templates, we used only AlphaFold2 models 1 and 2 since they utilize templates and the MSA data, while models 3, 4, and 5 do not use templates (Jumper, Evans et al. 2021) which was done by commenting out models 3-5 (lines 39-61 in /alphafold/model/config.py):

```
MODEL_PRESETS = {
    'monomer': (
        'model_1',
        'model_2',
  #      'model_3',
  #      'model_4',
  #      'model_5',
    ),
    'monomer_ptm': (
        'model_1_ptm',
        'model_2_ptm',
 #       'model_3_ptm',
 #       'model_4_ptm',
 #       'model_5_ptm',
    ),
    'multimer': (
        'model_1_multimer_v3',
        'model_2_multimer_v3',
#        'model_3_multimer_v3',
#        'model_4_multimer_v3',
```

```
#        'model_5_multimer_v3',
    ),
}
```

We ran AlphaFold2 with specific sequence data sets by replacing ./uniref90/uniref90.py with our sequence sets: Uniref90, Ortholog, Family for the MSA building step and specific template data sets to predict protein structures. Template implementation consisted of two parts: .cif files of the structures in `./pdb_mmcif/mmcif_files` and their sequence data in pdb70 files in `./pdb70` folder, they correspondingly need to be changed for AF2 to use specific sets of templates: PDB70, ActivePDB, ActiveAF2. The output was the 3D coordinates of amino acid residues, accompanied by a per-residue confidence score (pLDDT) that indicates the model's certainty regarding atoms in the neighborhood of each residue's prediction (Mariani, Biasini et al. 2013).

We introduced a variable `MSAlimit` that controls the number of sequences in the multiple sequence alignment used by AF2 model building by modifying the class `DataPipeline` (in /alphafold/data/pipeline.py). When AF2 has too many sequences in the MSA, it tends to ignore any templates provided to it. We also disabled other sequence databases like mgnify, bfd, small bfd, uniref30:

```
def __init__(self,
             jackhmmer_binary_path: str,
             hhblits_binary_path: str,
             uniref90_database_path: str,
             #mgnify_database_path: str,
             #bfd_database_path: Optional[str],
             #uniref30_database_path: Optional[str],
             #small_bfd_database_path: Optional[str],
             template_searcher: TemplateSearcher,
             template_featurizer: templates.TemplateHitFfeaturizer,
             #use_small_bfd: bool,
             #mgnify_max_hits: int = 501,
             uniref_max_hits: int = 10000,
             use_precomputed_msas: bool = False):
```

Benchmarking

The predicted structures were validated by comparing them to the benchmark PDB structures of kinases. The validation process relied on the pLDDT score and Root Mean Square Deviation (RMSD), measuring the average distance between atoms in the predicted and known structures. Model structures were aligned to benchmark structures with the program CE (Shindyalov and Bourne 1998) as implemented in PyMOL. The alignment was performed on the C-terminal domain of each structure. RMSD was measured for the activation loop backbone atoms (N, CA, C, O) after superposition of the C-terminal domains.

Two benchmarks were constructed. One contained substrate-bound structures from Table 1 with complete coordinates for the activation loop (22 kinases). The other consisted of 170 structures of 130

with complete activation loops that passed our active criteria from the PDB. For some kinases, there were multiple conformations that passed our criteria. We labeled the structure that most closely resembled substrate-bound structures as "conf1" with the others labeled "conf2,", "conf3," etc.

Data Availability and Reproducibility

To ensure the reproducibility of our study, all data including input sequences, predicted structures, and AlphaFold2 running scripts are accessible at http://dunbrack.fccc.edu/kincore/activemodels.

ACKNOWLEDGMENTS

This work was funded by NIH Grant R35 GM122517 (to RLD) and P30 CA006927 (to Fox Chase Cancer Center).

REFERENCES

Cheng, W., K. R. Munkvold, H. Gao, J. Mathieu, S. Schwizer, S. Wang, Y.-b. Yan, J. Wang, G. B. Martin and J. Chai (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**(6): 616-626.

Cohen, P., D. Cross and P. A. Jänne (2021). "Kinase drug discovery 20 years after imatinib: progress and future directions." Nature reviews drug discovery **20**(7): 551-569.

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

Derewenda, Z. S., L. Lee and U. Derewenda (1995). "The occurence of C–H··· O hydrogen bonds in proteins." Journal of molecular biology **252**(2): 248-262.

Fajer, M., Y. Meng and B. Roux (2017). "The activation of c-Src tyrosine kinase: conformational transition pathway and free energy landscape." The Journal of Physical Chemistry B **121**(15): 3352-3363.

Foda, Z. H., Y. Shan, E. T. Kim, D. E. Shaw and M. A. Seeliger (2015). "A dynamically coupled allosteric network underlies binding cooperativity in Src kinase." Nature communications **6**(1): 5939.

Frankish, A., S. Carbonell-Sala, M. Diekhans, I. Jungreis, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, C. Arnan and I. Barnes (2023). "GENCODE: reference annotation for the human and mouse genomes in 2023." Nucleic acids research **51**(D1): D942-D949.

Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics **28**(23): 3150-3152.

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

Heo, L. and M. Feig (2022). "Multi-state modeling of G-protein coupled receptors at experimental accuracy." Proteins: Structure, Function, and Bioinformatics **90**(11): 1873-1885.

Jacobs, M. D., P. R. Caron and B. J. Hare (2008). "Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: structure of LCK/imatinib complex." Proteins **70**(4): 1451-1460.

Joshi, M. K., R. A. Burton, H. Wu, A. M. Lipchik, B. P. Craddock, H. Mo, L. L. Parker, W. T. Miller and C. B. Post (2020). "Substrate binding to Src: A new perspective on tyrosine kinase substrate recognition from NMR and molecular dynamics." Protein Science **29**(2): 350-359.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.

Kanev, G. K., C. de Graaf, B. A. Westerman, I. J. de Esch and A. J. Kooistra (2021). "KLIFS: an overhaul after the first 5 years of supporting kinase research." Nucleic Acids Research **49**(D1): D562-D569.

Katso, R., R. Russell and T. Ganesan (1999). "Functional analysis of H-Ryk, an atypical member of the receptor tyrosine kinase family." Molecular and cellular biology **19**(9): 6427-6440.

Knighton, D. R., J. H. Zheng, L. F. Ten Eyck, V. A. Ashford, N.-H. Xuong, S. S. Taylor and J. M. Sowadski (1991). "Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase." Science **253**(5018): 407-414.

Kornev, A. P. and S. S. Taylor (2010). "Defining the conserved internal architecture of a protein kinase." Biochim Biophys Acta **1804**(3): 440-444.

Levinson, N. M., O. Kuchment, K. Shen, M. A. Young, M. Koldobskiy, M. Karplus, P. A. Cole and J. Kuriyan (2006). "A Src-like inactive conformation in the abl tyrosine kinase domain." PLoS biology **4**(5): e144.

Lin, K., J. Lin, W.-I. Wu, J. Ballard, B. B. Lee, S. L. Gloor, G. P. Vigers, T. H. Morales, L. S. Friedman and N. Skelton (2012). "An ATP-site on-off switch that restricts phosphatase accessibility of Akt." Science signaling **5**(223): ra37-ra37.

Luz, S., K. M. Cihil, D. L. Brautigan, M. D. Amaral, C. M. Farinha and A. Swiatecka-Urban (2014). "LMTK2-mediated phosphorylation regulates CFTR endocytosis in human airway epithelial cells." Journal of Biological Chemistry **289**(21): 15080-15093.

Mariani, V., M. Biasini, A. Barbato and T. Schwede (2013). "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests." Bioinformatics **29**(21): 2722-2728.

Modi, V. and R. Dunbrack (2022). "Kincore: a web resource for structural classification of protein kinases and their inhibitors." Nucleic Acids Research **50**(D1): D654-D664.

Modi, V. and R. L. Dunbrack (2019). "Defining a new nomenclature for the structures of active and inactive kinases." Proceedings of the National Academy of Sciences **116**(14): 6818-6827.

Modi, V. and R. L. Dunbrack, Jr. (2019). "A Structurally Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains." Sci Rep **9**(1): 19790.

Schindler, T., W. Bornmann, P. Pellicena, W. T. Miller, B. Clarkson and J. Kuriyan (2000). "Structural mechanism for STI-571 inhibition of abelson tyrosine kinase." Science **289**(5486): 1938-1942.

Seal, R. L., B. Braschi, K. Gray, T. E. Jones, S. Tweedie, L. Haim-Vilmovsky and E. A. Bruford (2023). "Genenames. org: the HGNC resources in 2023." Nucleic Acids Research **51**(D1): D1003-D1009.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." Protein Engineering **11**(9): 739-747.
Steinegger, M., M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger and J. Söding (2019). "HH-suite3 for fast remote homology detection and deep protein annotation." BMC bioinformatics **20**(1): 1-15.

Suijkerbuijk, S. J., T. J. van Dam, G. E. Karagöz, E. von Castelmur, N. C. Hubner, A. M. Duarte, M. Vleugel, A. Perrakis, S. G. Rüdiger and B. Snel (2012). "The vertebrate mitotic checkpoint protein BUBR1 is an unusual pseudokinase." Developmental cell **22**(6): 1321-1329.

Ung, P. M.-U., R. Rahman and A. Schlessinger (2018). "Redefining the protein kinase conformational space with machine learning." Cell chemical biology **25**(7): 916-924. e912.

Varadi, M., S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood and A. Laydon (2022). "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." Nucleic acids research **50**(D1): D439-D444.

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

Yang, J., J. Wu, J. M. Steichen, A. P. Kornev, M. S. Deal, S. Li, B. Sankaran, V. L. Woods Jr and S. S. Taylor (2012). "A conserved Glu–Arg salt bridge connects coevolved motifs that define the eukaryotic protein kinase fold." Journal of molecular biology **415**(4): 666-679.

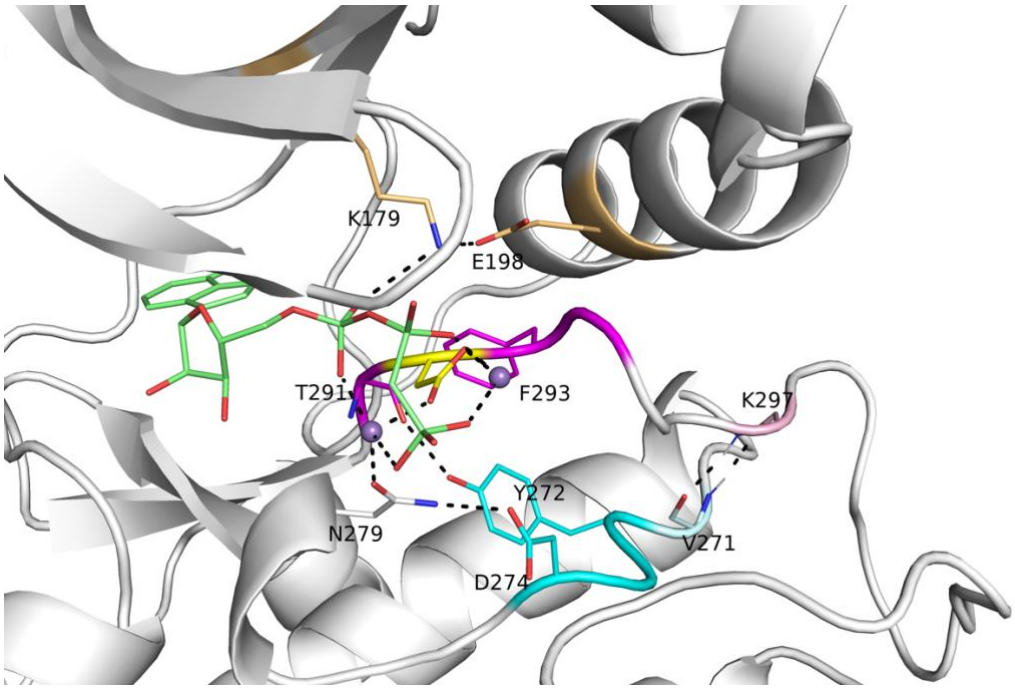


Figure 1. Active site of human AKT1 (PDB:4ekk, chain A). Residues making hydrogen bonding interactions with ATP (green sticks), Mg^{2+} (purple spheres), the catalytic aspartic acid residue of the HRD motif (in AKT1, this is YRD; residues 272-274, cyan), and the aspartic residue (yellow) of the XDFG 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.

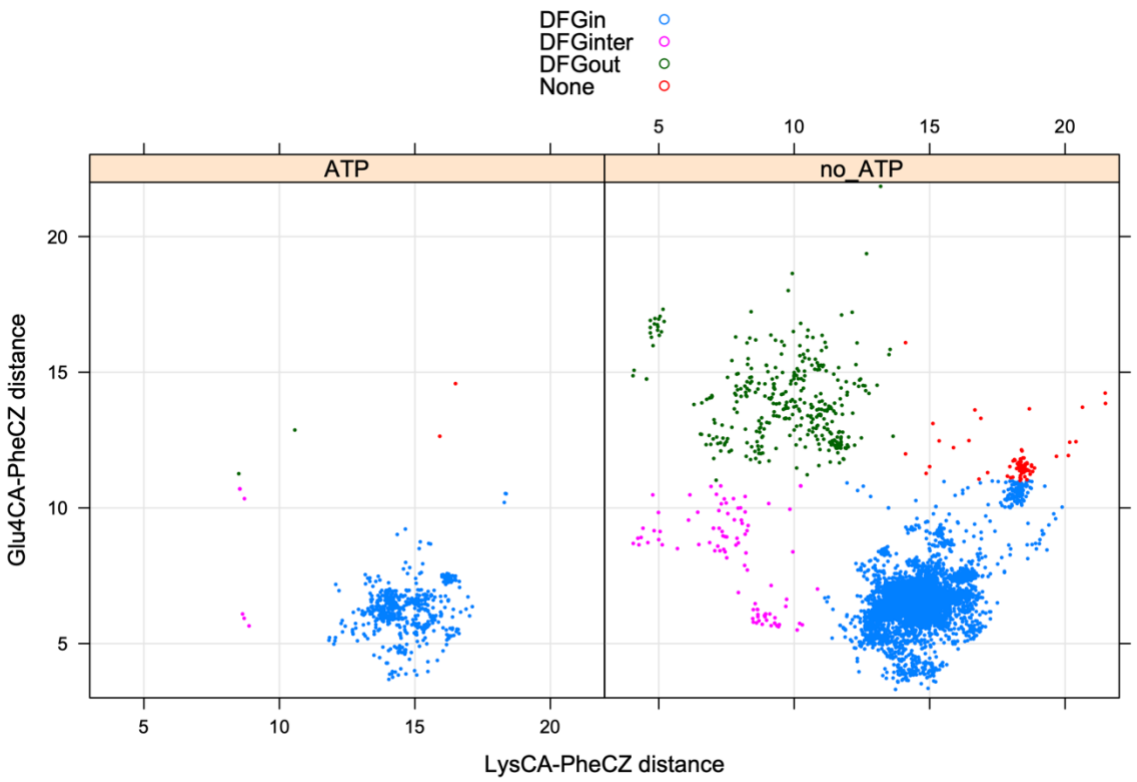


Figure 2. Defining DFGin and DFGout structures. Distances of the Cζ atom of DFG-Phe from the Cα atoms of the salt bridge Lys residue and Glu4 (4 residues after the salt bridge Glu).

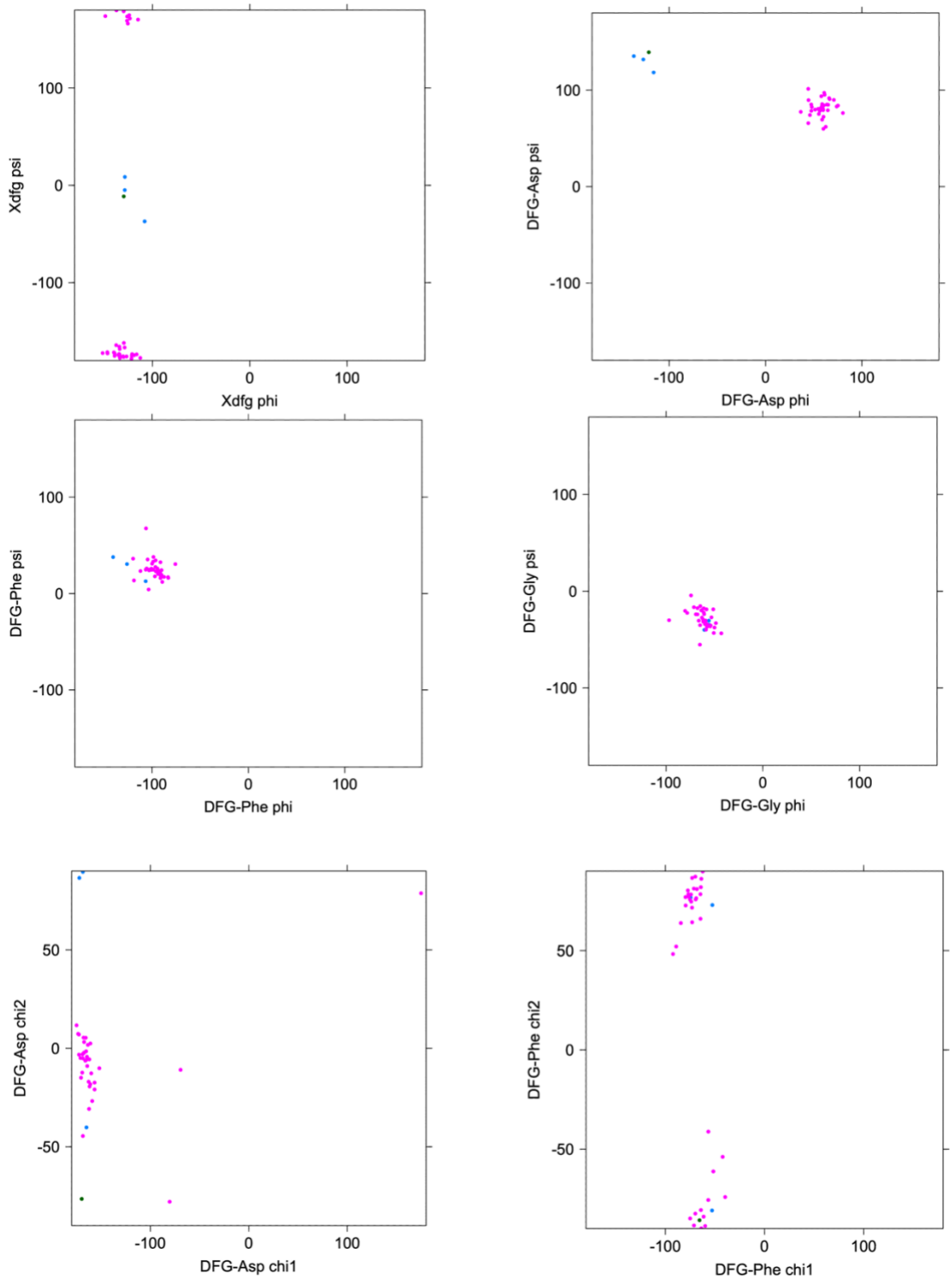


Figure 3. Ramachandran plots and side-chain dihedral angle plots for the XDFG motif residues in substrate-bound structures in Table 1. BLAminus structures are shown in magenta and ABAminus structures are shown in blue.

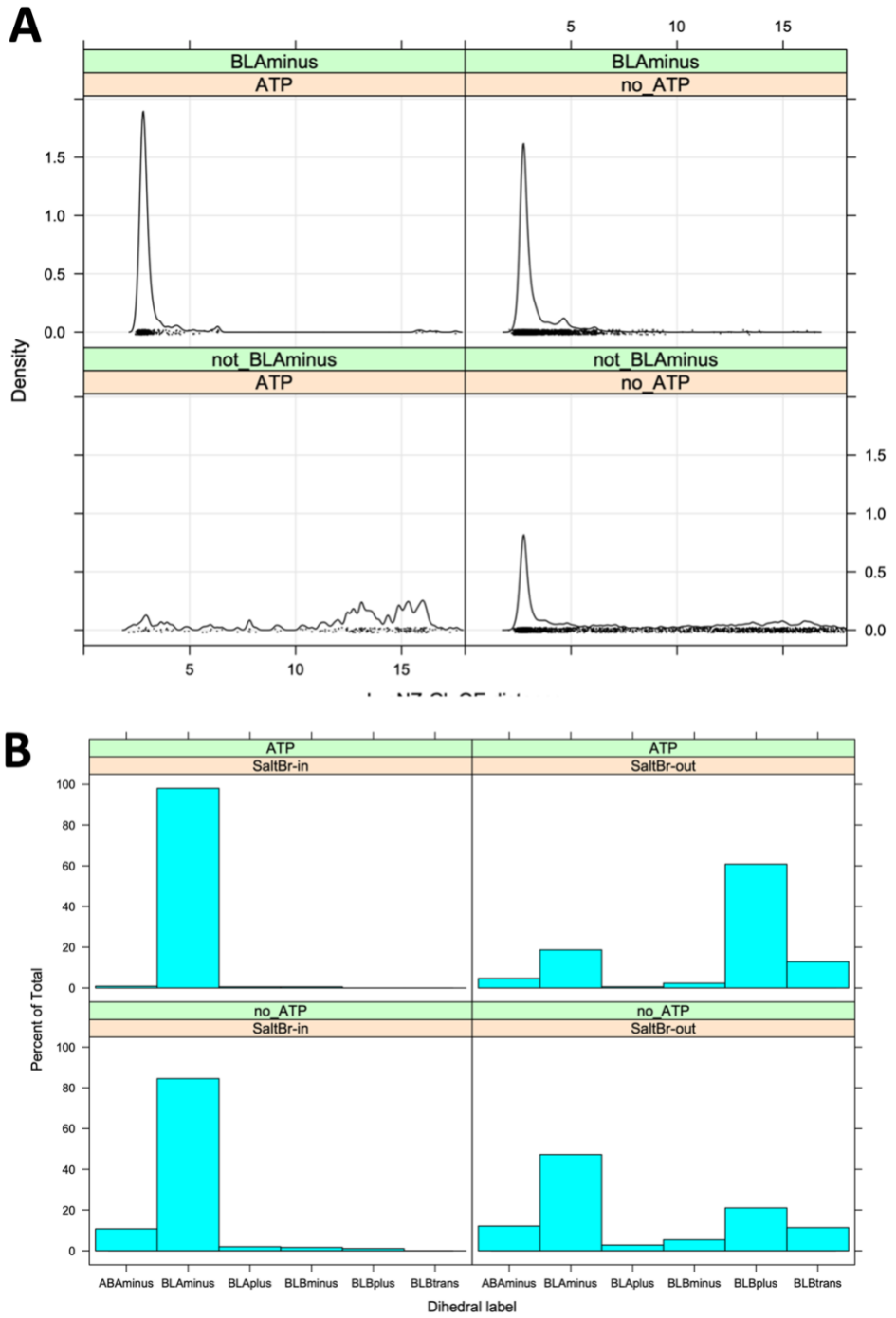


Figure 4. Relationship of the XDFG motif dihedral angle conformational state and the N-termainal domain salt bridge.

A. Minimum distance of the Nζ atom from the Lys residue and the Oε1 or Oε2 atoms of the Glu residue of the salt bridge in structures with or without ATP and in or out of the BLAminus conformation.

B. Distribution of dihedral angle states in ATP-bound and ATP-unbound structures in the presence of the absence of the N-terminal domain salt bridge (minimum Nζ/Oε distance cutoff 3.6 Å). When the salt bridge and ATP are present, 99% of structures are in the BLAminus conformation for the XDF motif.

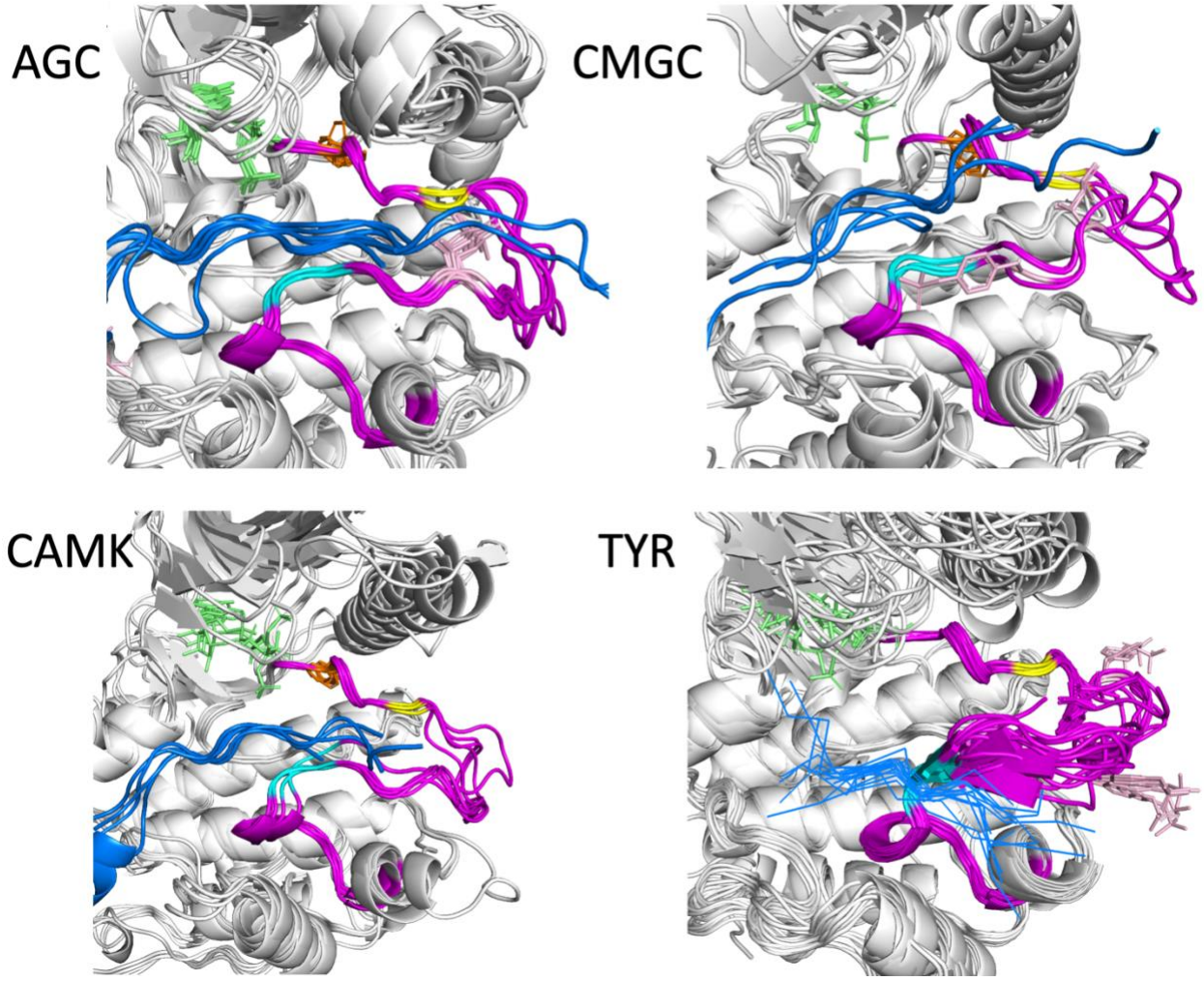


Figure 5. Substrate-bound structures in the AGC, CAMK, CMGC, and TYR families from Table 1. In each figure, the substrate peptides (or piece of longer protein) are in blue and the activation loop is in magenta. ATP or any analogue is shown in green sticks. The Phe of the DFG motif is shown in orange sticks and phosphorylated residues in the activation loop are in pink. The sixth residue of the activation loop (DFGxxX) is in yellow, while the 8th and 9th residues from the end of the activation loop are in cyan (XXxxxxAPE).

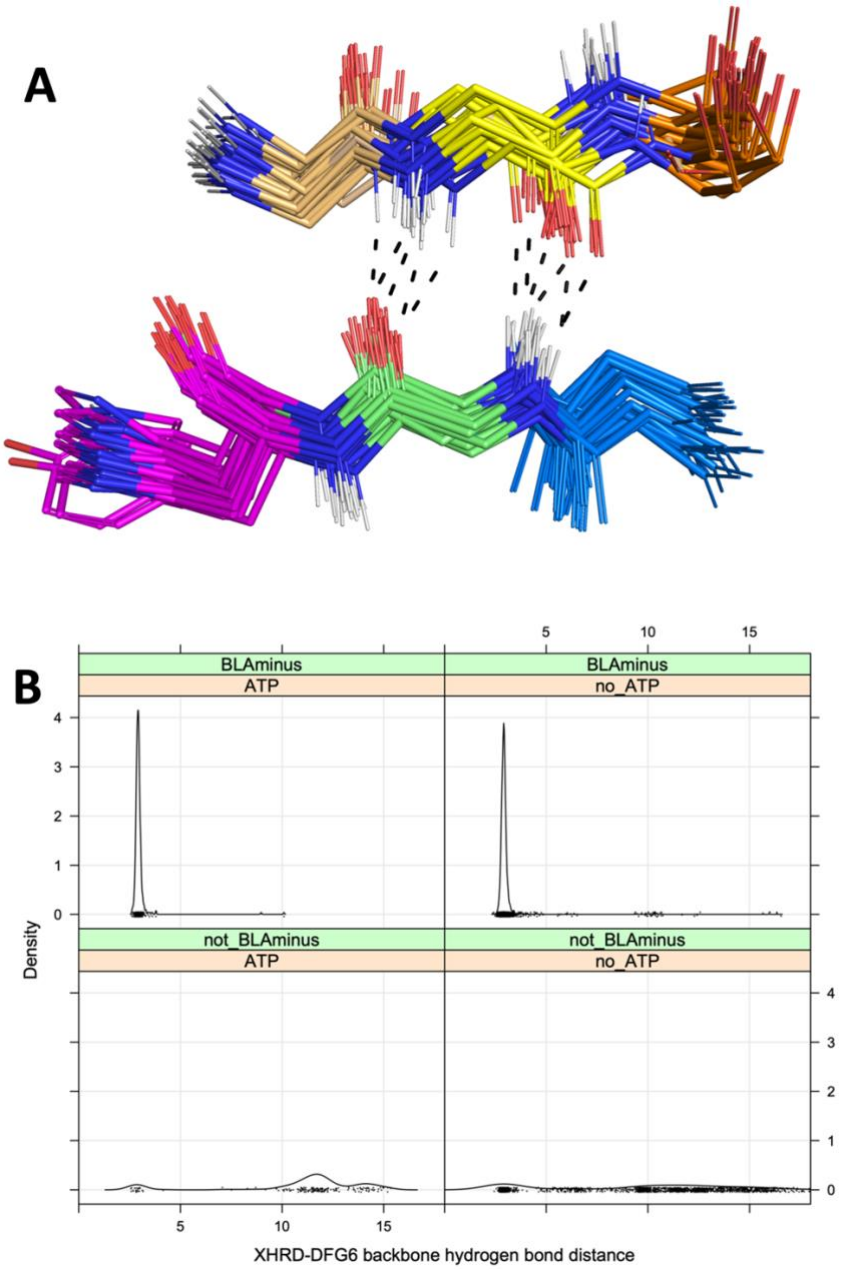


Figure 6. Interactions of residue 6 of the activation loop and the residue before the HRD motif ("XHRD")

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. Distribution of the XHRD-DFG6 backbone-backbone hydrogen bond distance of ATP bound and unbound structures in the BLAminus and other states. 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 is any amino acid.

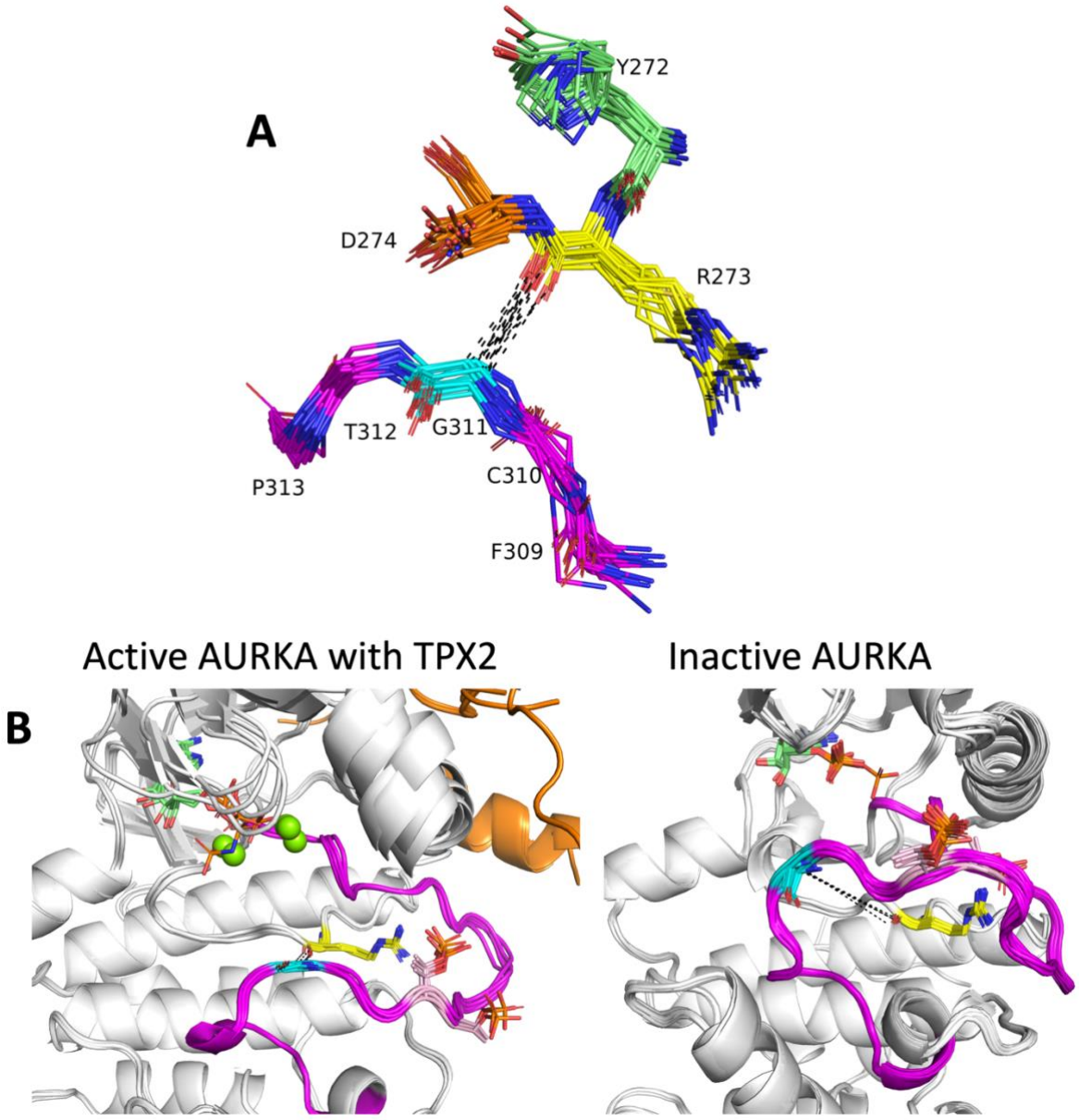


Figure 7. Role of the C-terminal segment of the activation loop in substrate binding.

A. Contact between Ca of APE9 residue (XxxxxxAPE) and backbone carbonyl oxygen of Arg residue of HRD motif. His/Tyr (green), Arg (yellow), and Asp (orange) of HRD/YRD motif 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. Structures from the AGC, CAMK, CMGC, STE, and TKL kinases in Table 1 are shown. B. Two conformations of human AURKA. A. Active structures with bound TPX2 (orange): PDB: 1ol5, 3e5a, 3ha6, 5lxm, 6vph.

B. Inactive structures without TPX2. PDB: 4dee, 5dt3, 5oro, 5oso, 6i2u, 6r49, 6r4d and others. Gly291 Ca (cyan sticks) is in contact with 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 in Figure 5, while the inactive structures would block substrate binding.

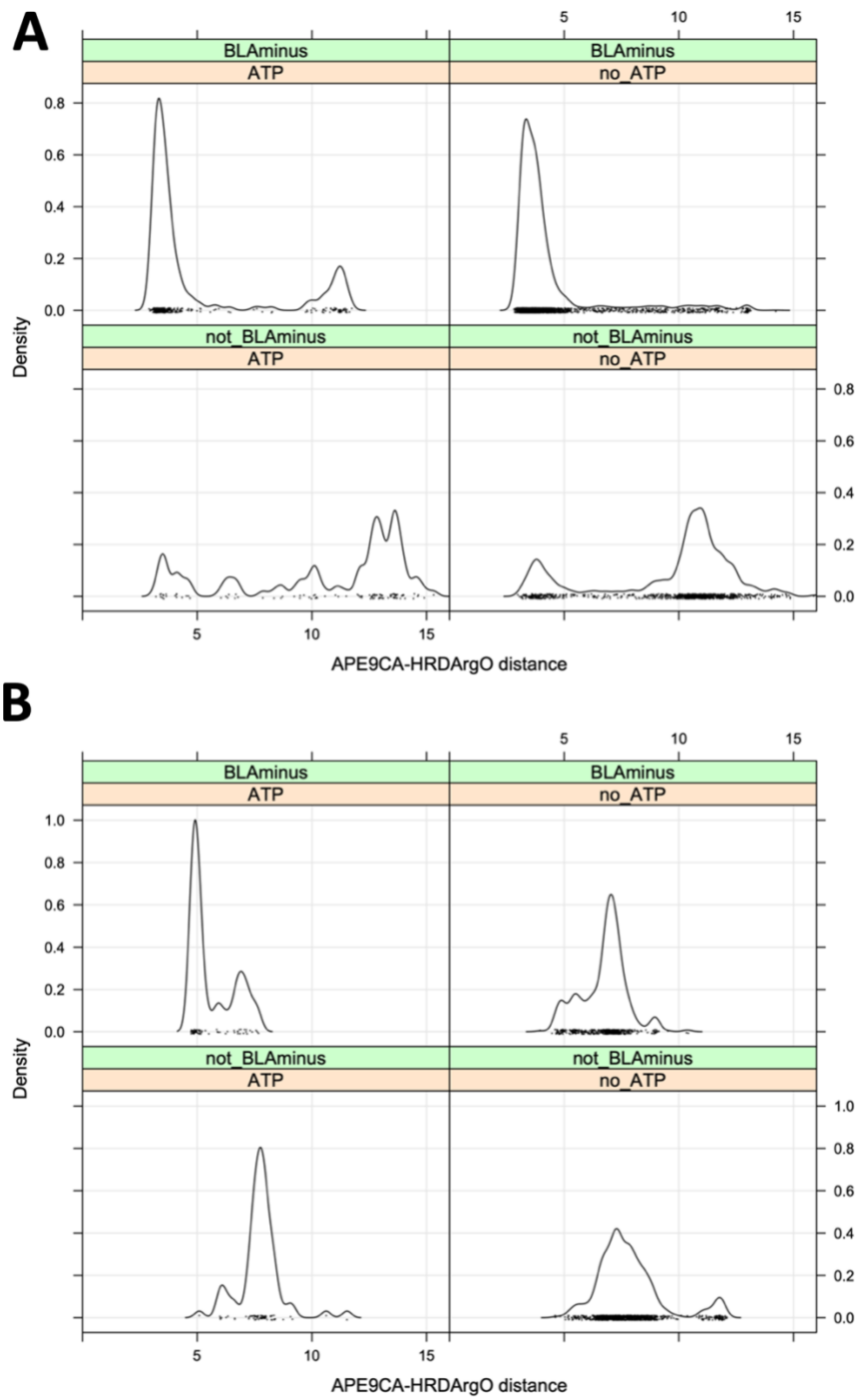


Figure 8.Criterion of residue 9 from the end of the activation loop ("APE9")

A. Distribution of the APE9-CA/hRd-O distance n ATP-bound and ATP-unbound structures.

B. Distribution of the APE9-CA/hRd-O distance in ATP-bound and ATP-unbound tyrosine kinase structures. The peak at 5 Å in ATP-bound/BLAminus structures are all structures of FGFR2. The slightly longer distances from up to 8 Å are more characteristic of active tyrosine kinase structures. Longer distances (over 10 Å) are observed in non-BLAminus/non-ATP-bound structures (lower right panel).

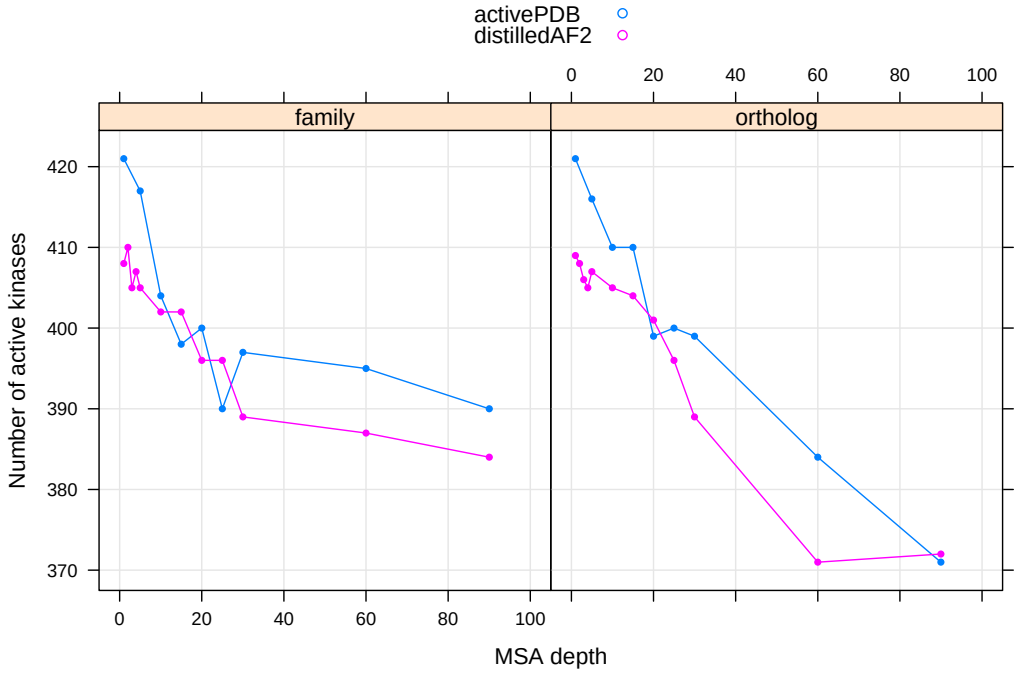


Figure 9. Number of catalytic kinases with active models produced by different template data sources (activePDB and distilledAF2) and different sequence sources (family kinases and orthologous kinases for each target).

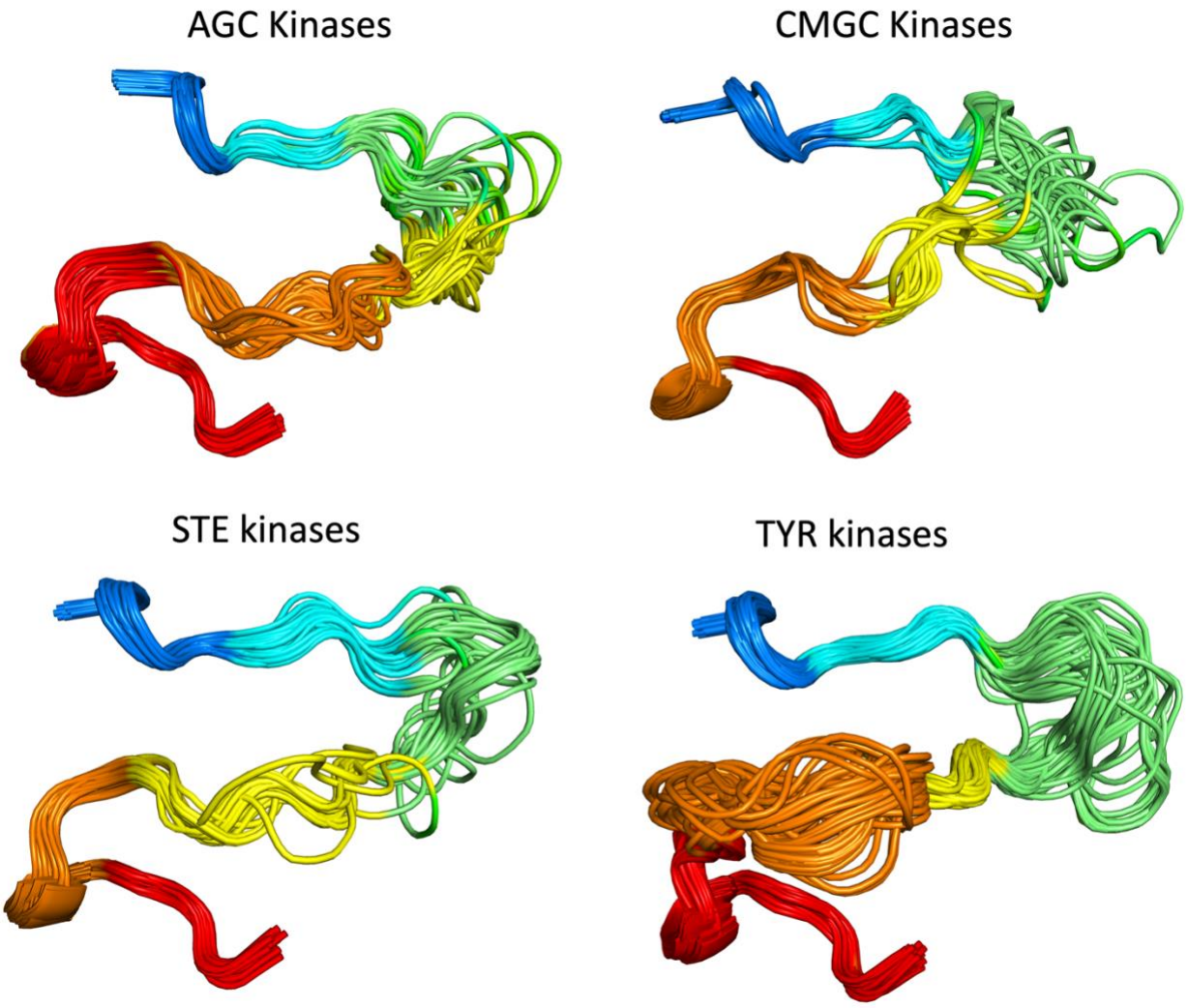


Figure 10. Examples of AlphaFold2 models of the active forms of 51 AGC kinases, 65 CMGC kinases, 31 STE kinases, and 77 TYR kinase.
For clarity, some structures with long disordered regions within the activation loop are not shown in each family.

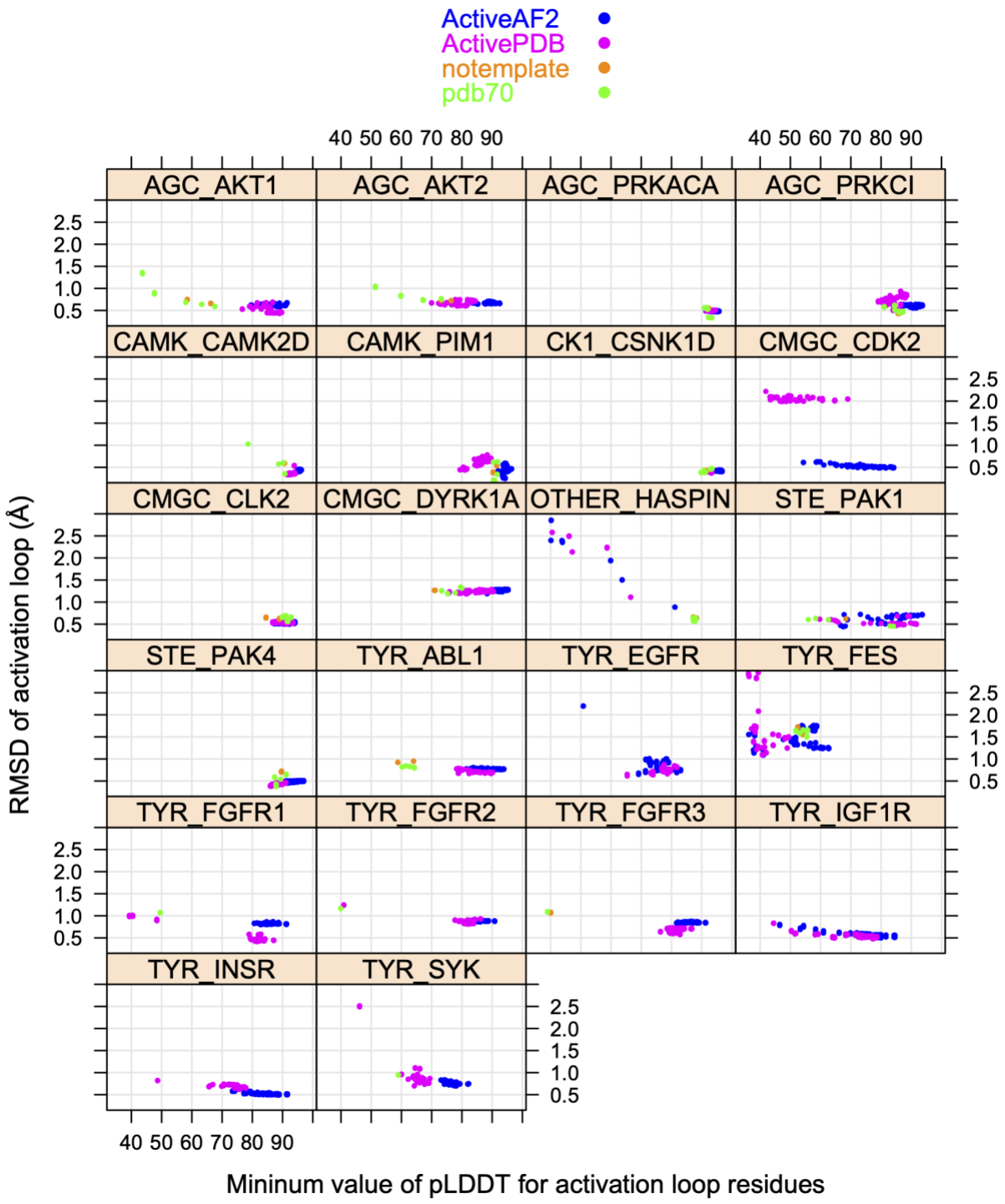


Figure 11. RMSD values for 22 substrate-bound structures from Table 1 versus the minimum value of pLDDT across the activation loop of each model. Models from different template data sets are shown in different colors: active structures from the PDB ("ActivePDB"), active models from AlphaFold2 ("ActiveAF2"), no templates provided to AF2 ("notemplate"), and AlphaFold2's default template database ("PDB70").

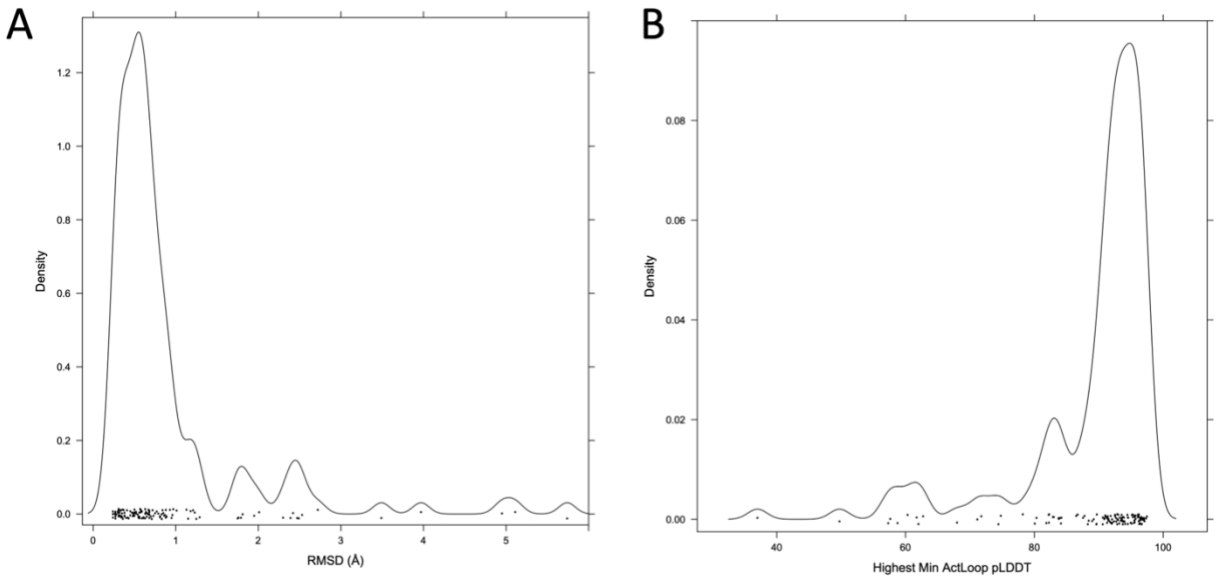


Figure 12. Benchmark of 130 active kinases in the PDB with full coordinates for the activation loop.

A. Distribution of RMSD values (with kernel density estimate) for the top scoring model (minimum value of the pLDDT across the activation loop).

B. Distribution of min_pLDDT values for these 130 models.

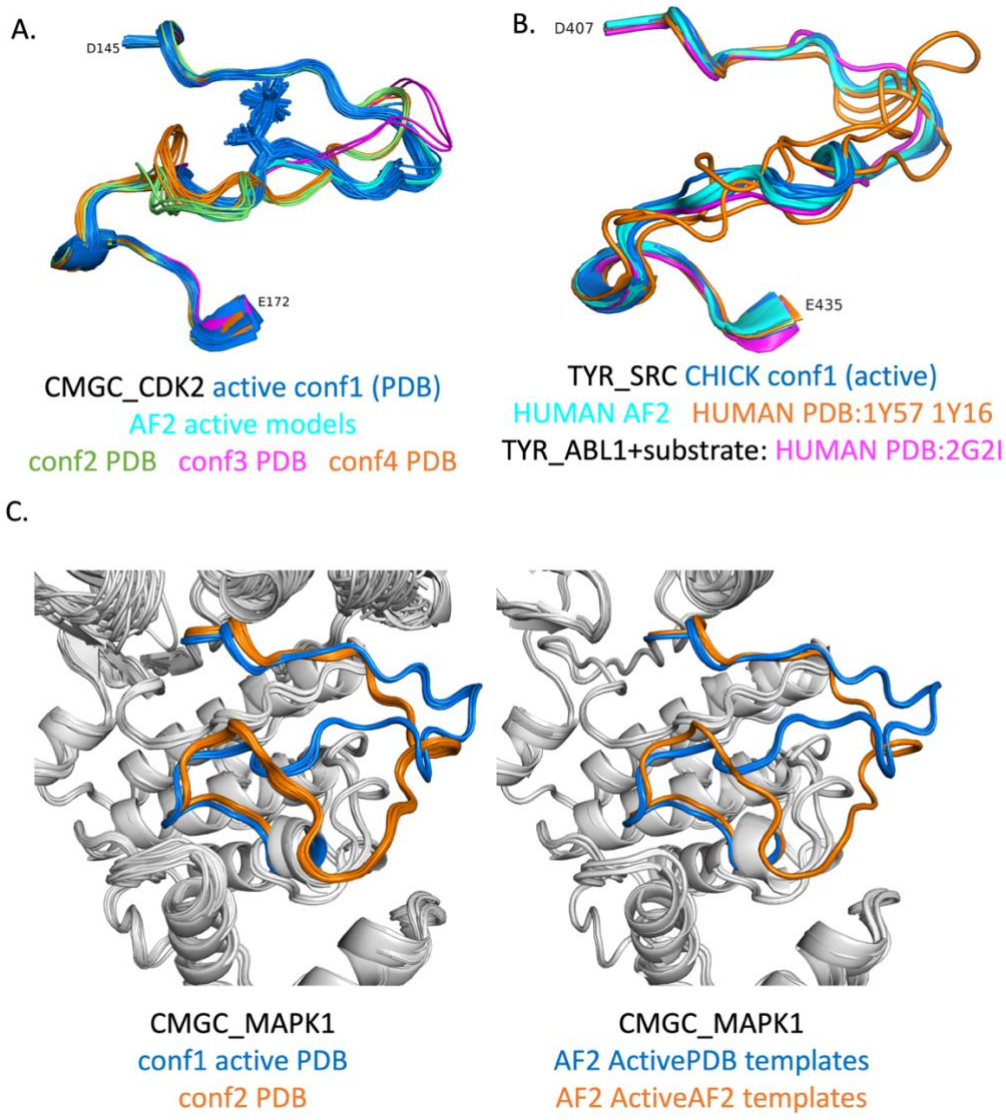


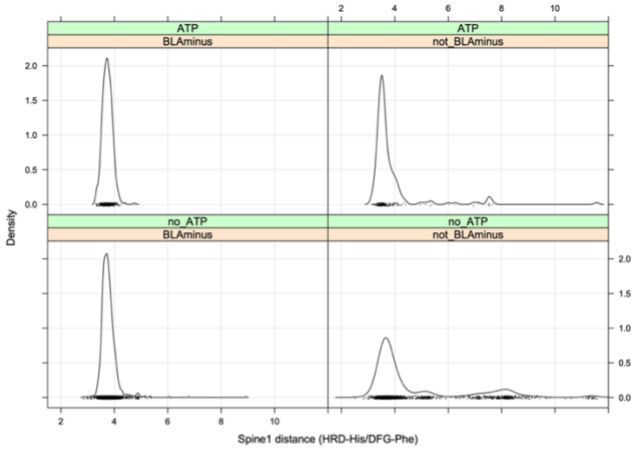
Figure 13. Structures of activation loops from the benchmark and corresponding AF2 models.

A. CMGC_CDK2 has four dominant conformations among structures that pass our "Active" criteria in the PDB. The Conformation 1 cluster contains the substrate-bound structures listed in Table 1, and is also the only cluster that contains phosphorylated activation loops. The AlphaFold2 models most closely resemble Conformation 1.

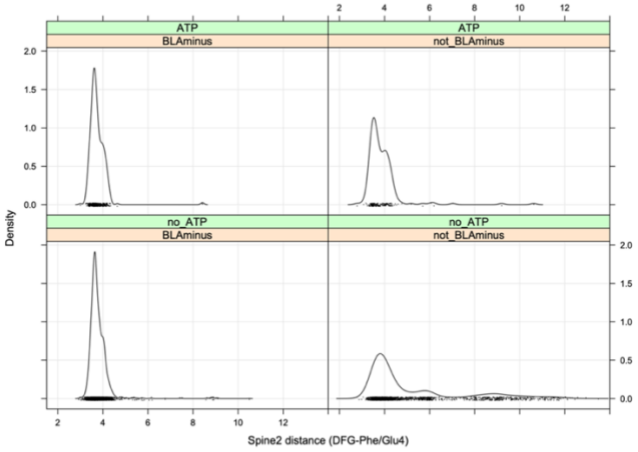
B.For TYR_SRC, we used a chicken SRC structure (PDB: 3DQW) as the benchmark structure since it most closely resembled substrate0bound structures of other TYR kinases such as ABL1 (e.g., PDB:2G2I in Table 1). The human structures which pass our criteria (PDB: 1Y57 and 1Y16 in orange) are quite different in much of the activation loop away from the first few and last few residues of the DFG...APE sequence. PDB: 1Y57 is often used as the basis of molecular dynamics simulations of "Active SRC" even though it is not likely the substrate-binding conformation.

C. CMGC_MAPK1 has two dominant conformations in the PDB (left panel), one of which resembles substrate-bound structures in Table 1 ("conf1") while the other has a bulge towards the C-terminus of the activation loop. AlphaFold2 reproduces both conformations (right panel), conf1 from ActivePDB templates and conf2 from ActiveAF2 ("distillation") templates.

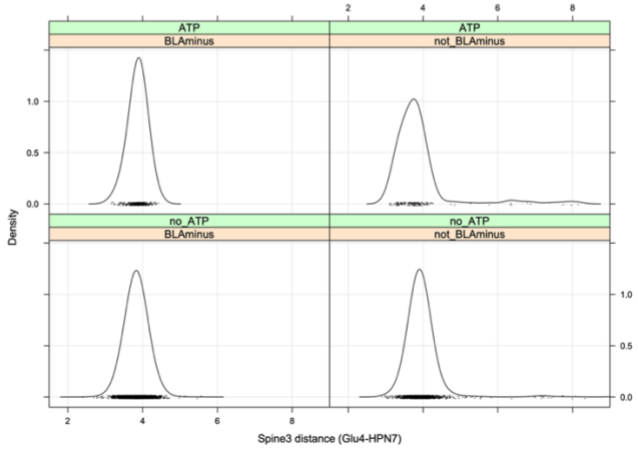
A.



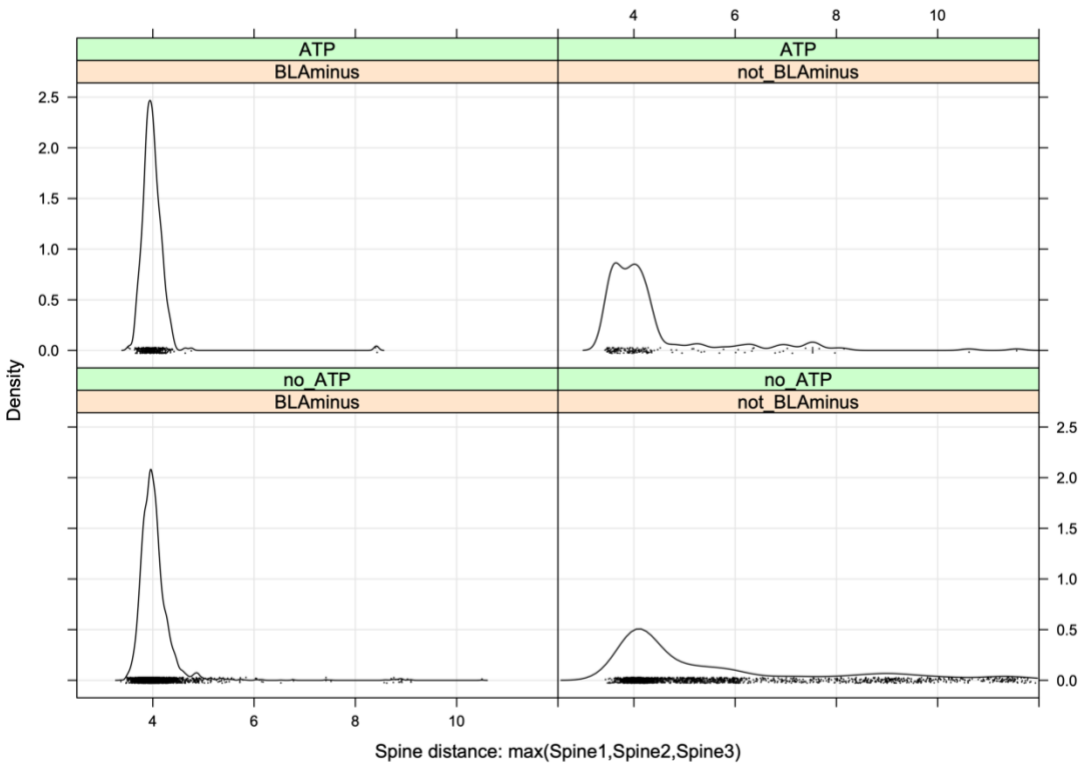
B.



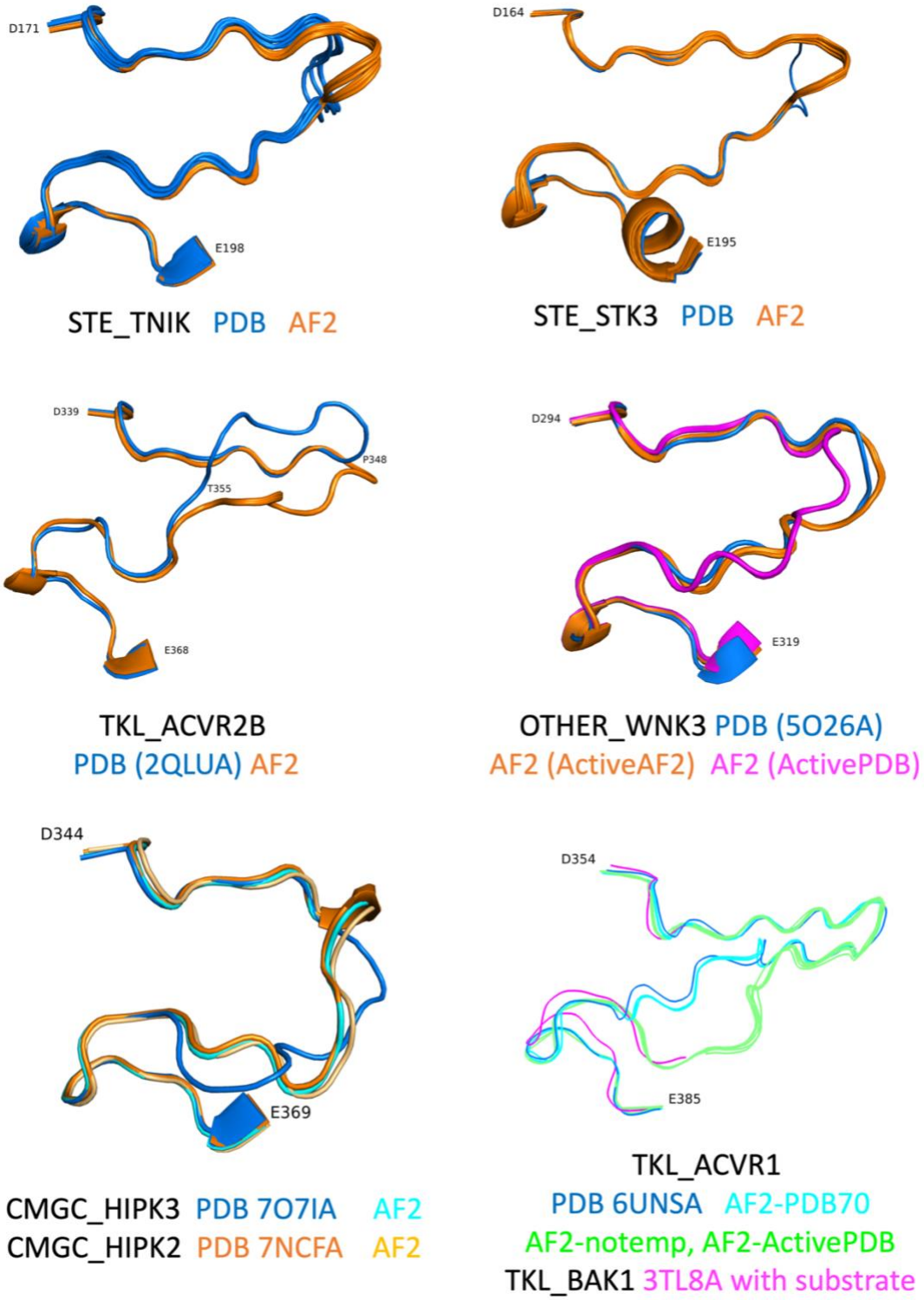
C.



Supplementary Figure 1. Distribution of the Spine1, Spine2, and Spine 3 distances. The Spine distances are defined as the closest distance among all side-chain atom pairs between the two residues.

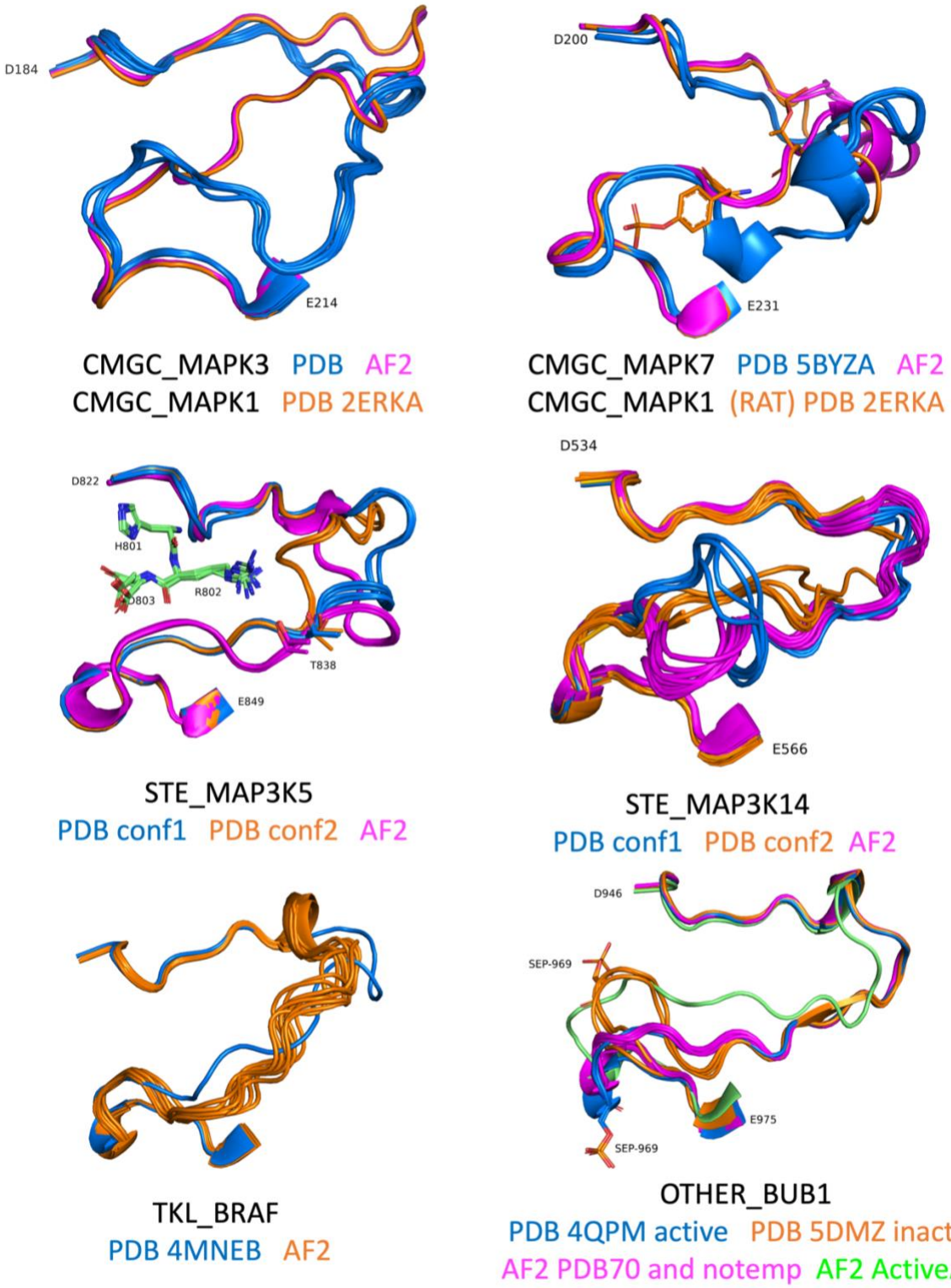


Supplementary Figure 2. Distribution of the Spine distance, which is the maximum of Spine1, Spine2, Spine3 (Supplementary Figure 1) for each kinase structure.



Supplementary Figure 3. Benchmark structures with large RMSD to the best scoring AlphaFold2 models. STE_TNIK, STE_STK2, and TKL_ACVR2B show deviations between the PDB structures (blue) and AlphaFold2 models (orange) only in the outer regions of the activation loop. For OTHER_WNK3, the highest scoring pLDDT model is from an ActivePDB template (magenta) while an ActiveAF template model is much closer (orange) to the benchmark structure (PDB:5O26A, blue). For CMGC_HIPK3, the AF2 model more closely resembles an active PDB structure of

HIPK2 (PDB: 7NCFA, orange) as well as an AF2 model of HIPK2 (light orange) than a PDB structure of HIPK3 itself (7O7IA, blue). The activation loop sequences are identical in 25 of 26 positions, and it is likely that AF2 is correct in modeling very similar conformations of the activation loop. The PDB structure of TKL_ACVR1 (PDB: 6UNSA. blue) resembles the AF2 model made from PDB70 (cyan), while the ActivePDB AF2 model and the no-template AF2 model more closely resembles a substrate-bound PDB structure of TKL_BAK1 from *Arabidopsis* (PDB: 3TL8, magenta, substrate not shown). It is likely that the ActivePDB AF2 model (green) is a substrate-binding structure, while the PDB structure is not.



Supplementary Figure 4. Benchmark structures with large RMSD to the best scoring AlphaFold2 models. CMGC_MAPK3 and CMGC_MAPK7 show the same bulge toward the C-terminal end of the activation loop that is present in CMGC_MAPK1 (Figure 13C, main paper) in PDB structures (blue). The AF2 structures (magenta) more closely resemble RAT CMGC_MAPK1 (PDB:2ERKA, orange) shown in both figures. It is likely the AF2 structures are correct substrate-binding forms of MAPK3 and MAPK7. For STE_MAP3K5, the two dominant PDB conformations

(conf1 and conf2) are not typical of substrate-binding structures of STE kinases in the PDB. They also differ substantially from each other in the outer portions of the activation loop. The AF2 models (magenta) are closer to substrate-binding structures of STE kinases in Table 1 (PDB:2q0nA, 4zy4A). For STE_MAP3K14, the AF2 structures are quite different than either of the dominant PDB conformations for reasons that are unknown. TKL_BRAF is not that close to other kinases in the PDB other than RAF1, and the AF2 models are quite different than the benchmark one active structure with a complete activation loop in the PDB (4MNEB). OTHER_BUB1 is also distantly related to all other kinases; in the PDB there are two conformations, one of which has a large bulge blocking the substrate binding site (orange). Most of these structures are not phosphorylated on Ser969, although one of them is (shown in the figure). The other PDB structures (PDB:4QPM, blue) is more likely to be substrate-capable and most of these structures are phosphorylated on Ser969. The ActiveAF2-template-based model is quite different from both PDB conformations, while the PDB70 and no-template AF2 models (magenta) are much closer to the Active PDB structure (4QPM) and are most likely correct, even though they do not have pLDDT values as high as the ActiveAF2-template structures (green).

Supplementary Table 1. Catalytic kinase domains in the human proteome

N	N (Fam)	Fam_Gene	SwissProt ID	Gene	Uniprot Acc.	Kinase Start	Kinase End	Kinase Length	Length protein
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_PRPF4B	PRP4B_HUMAN	PRPF4B	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_NEK10	NEK10_HUMAN	NEK10	Q6ZWH5	511	793	283	1172
222	3	NEK_NEK11	NEK11_HUMAN	NEK11	Q8NG66	21	295	275	645
223	4	NEK_NEK2	NEK2_HUMAN	NEK2	P51955	1	279	279	445
224	5	NEK_NEK3	NEK3_HUMAN	NEK3	P51956	1	265	265	506
225	6	NEK_NEK4	NEK4_HUMAN	NEK4	P51957	1	269	269	841
226	7	NEK_NEK5	NEK5_HUMAN	NEK5	Q6P3R8	1	267	267	708
227	8	NEK_NEK6	NEK6_HUMAN	NEK6	Q9HC98	37	313	277	313
228	9	NEK_NEK7	NEK7_HUMAN	NEK7	Q8TDX7	26	302	277	302
229	10	NEK_NEK8	NEK8_HUMAN	NEK8	Q86SG6	1	266	266	692
230	11	NEK_NEK9	NEK9_HUMAN	NEK9	Q8TD19	44	316	273	979

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	70	TYR_RYK	RYK_HUMAN	RYK	P34925	322	606	285	607
426	71	TYR_SRC	SRC_HUMAN	SRC	P12931	262	529	268	536
427	72	TYR_SRMS	SRMS_HUMAN	SRMS	Q9H3Y6	222	488	267	488
428	73	TYR_SYK	KSYK_HUMAN	SYK	P43405	362	635	274	635
429	74	TYR_TEC	TEC_HUMAN	TEC	P42680	362	629	268	631
430	75	TYR_TEK	TIE2_HUMAN	TEK	Q02763	816	1102	287	1124
431	76	TYR_TIE1	TIE1_HUMAN	TIE1	P35590	831	1117	287	1138
432	77	TYR_TNK1	TNK1_HUMAN	TNK1	Q13470	108	387	280	666
433	78	TYR_TNK2	ACK1_HUMAN	TNK2	Q07912	118	395	278	1038
434	79	TYR_TXK	TXK_HUMAN	TXK	P42681	263	527	265	527
435	80	TYR_TYK2-2	TYK2_HUMAN	TYK2	P29597	889	1179	291	1187
436	81	TYR_TYRO3	TYRO3_HUMAN	TYRO3	Q06418	510	796	287	890
437	82	TYR_YES1	YES_HUMAN	YES1	P07947	269	536	268	543
438	83	TYR_ZAP70	ZAP70_HUMAN	ZAP70	P43403	329	603	275	619

Supplementary Table 2. Pseudokinase domains in the human proteome

N	N (Fam)	Fam_Gene	SwissProt ID	Gene	Uniprot Acc.	Kinase Start	Kinase End	Kinase Length	Length protein
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_PLK5	PLK5_HUMAN	PLK5	Q496M5	1	74	74	336
4	4	CAMK_PSKH2	KPSH2_HUMAN	PSKH2	Q96QS6	55	328	274	385
5	5	CAMK_STK40	STK40_HUMAN	STK40	Q8N2I9	27	338	312	435
6	6	CAMK_TRIB1	TRIB1_HUMAN	TRIB1	Q96RU8	86	346	261	372
7	7	CAMK_TRIB2	TRIB2_HUMAN	TRIB2	Q92519	56	316	261	343
8	8	CAMK_TRIB3	TRIB3_HUMAN	TRIB3	Q96RU7	63	323	261	358
9	9	CAMK_TTN	TITIN_HUMAN	TTN	Q8WZ42	32170	32440	271	34350
10	1	CK1_VRK3	VRK3_HUMAN	VRK3	Q8IV63	158	463	306	474
11	1	OTHER_BUB1B	BUB1B_HUMAN	BUB1B	O60566	758	1029	272	1050
12	2	OTHER_EIF2AK4-1	E2AK4_HUMAN	EIF2AK4	Q9P2K8	272	547	276	1649
13	3	OTHER_MLKL	MLKL_HUMAN	MLKL	Q8NB16	193	471	279	471
14	4	OTHER_NRBP1	NRBP_HUMAN	NRBP1	Q9UHY1	56	335	280	535
15	5	OTHER_NRBP12	NRBP2_HUMAN	NRBP2	Q9NSY0	29	314	286	501
16	6	OTHER_PAN3	PAN3_HUMAN	PAN3	Q58A45	472	759	288	887
17	7	OTHER_PEAK1	PEAK1_HUMAN	PEAK1	Q9H792	1319	1673	355	1746
18	8	OTHER_PEAK3	PEAK3_HUMAN	PEAK3	Q6ZS72	163	405	243	473
19	9	OTHER_PIK3R4	PI3R4_HUMAN	PIK3R4	Q99570	18	321	304	1358
20	10	OTHER_POMK	SG196_HUMAN	POMK	Q9H5K3	73	341	269	350
21	11	OTHER_PRAG1	PRAG1_HUMAN	PRAG1	Q86YV5	984	1335	352	1406
22	12	OTHER_PXK	PXK_HUMAN	PXK	Q7Z7A4	138	404	267	578
23	13	OTHER_RNASEL	RN5A_HUMAN	RNASEL	Q05823	353	594	242	741
24	14	OTHER_RPS6KC1	KS6C1_HUMAN	RPS6KC1	Q96S38	332	1066	735	1066
25	15	OTHER_RPS6KL1	RPKL1_HUMAN	RPS6KL1	Q9Y6S9	145	547	403	549
26	16	OTHER_SCYL1	SCYL1_HUMAN	SCYL1	Q96KG9	6	271	266	808
27	17	OTHER_SCYL2	SCYL2_HUMAN	SCYL2	Q6P3W7	24	335	312	929
28	18	OTHER_SCYL3	PACE1_HUMAN	SCYL3	Q8IZE3	3	253	251	742
29	19	OTHER_STK31	STK31_HUMAN	STK31	Q9BXU1	702	980	279	1019
30	20	OTHER_STKLD1	STKL1_HUMAN	STKLD1	Q8NE28	20	305	286	680
31	21	OTHER_TBCK	TBCK_HUMAN	TBCK	Q8TEA7	1	281	281	893
32	22	OTHER_TEX14	TEX14_HUMAN	TEX14	Q8IWB6	219	520	302	1497
33	23	OTHER_ULK4	ULK4_HUMAN	ULK4	Q96C45	1	288	288	1275
34	1	RGC_GUCY2C	GUC2C_HUMAN	GUCY2C	P25092	468	767	300	1073
35	2	RGC_GUCY2D	GUC2D_HUMAN	GUCY2D	Q02846	502	815	314	1103
36	3	RGC_GUCY2F	GUC2F_HUMAN	GUCY2F	P51841	505	853	349	1108
37	4	RGC_NPR1	ANPRA_HUMAN	NPR1	P16066	507	829	323	1061
38	5	RGC_NPR2	ANPRB_HUMAN	NPR2	P20594	491	814	324	1047
39	1	STE_STRADA	STRAA_HUMAN	STRADA	Q7RTN6	61	387	327	431
40	2	STE_STRADB	STRAB_HUMAN	STRADB	Q9C0K7	50	377	328	418
41	1	TKL_ILK	ILK_HUMAN	ILK	Q13418	185	452	268	452
42	2	TKL_IRAK2	IRAK2_HUMAN	IRAK2	O43187	197	509	313	625
43	3	TKL_IRAK3	IRAK3_HUMAN	IRAK3	Q9Y616	152	453	302	596

44	4	TKL_KSR1	KSR1_HUMAN	KSR1	Q8IVT5	605	887	283	923
45	5	TKL_KSR2	KSR2_HUMAN	KSR2	Q6VAB6	658	938	281	950
46	1	TYR_EPHA10	EPHAA_HUMAN	EPHA10	Q5JZY3	637	910	274	1008
47	2	TYR_EPHB6	EPHB6_HUMAN	EPHB6	O15197	662	925	264	1021
48	3	TYR_ERBB3	ERBB3_HUMAN	ERBB3	P21860	701	975	275	1342
49	4	TYR_JAK1-1	JAK1_HUMAN	JAK1	P23458	575	855	281	1154
50	5	TYR_JAK2-1	JAK2_HUMAN	JAK2	O60674	537	815	279	1132
51	6	TYR_JAK3-1	JAK3_HUMAN	JAK3	P52333	513	787	275	1124
52	7	TYR_PTK7	PTK7_HUMAN	PTK7	Q13308	788	1070	283	1070
53	8	TYR_ROR1	ROR1_HUMAN	ROR1	Q01973	465	756	292	937
54	9	TYR_ROR2	ROR2_HUMAN	ROR2	Q01974	465	756	292	943
55	10	TYR_RYK	RYK_HUMAN	RYK	P34925	322	606	285	607
56	11	TYR_STYK1	STYK1_HUMAN	STYK1	Q6J9G0	105	390	286	422
57	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. The RGC proteins contain active guanylyl cyclase domains; the kinase domains are inactive. PLK5 contains a truncated kinase domain at its N-terminus. The mouse ortholog contains a full kinase domain.