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.