Combining structural bioinformatics and deep-learning-based protein structure prediction:

AlphaFold2 models of all *437* catalytically competent human kinases in the active form

Bulat Faezov
Roland Dunbrack

FOX CHASE
CANCER CENTER

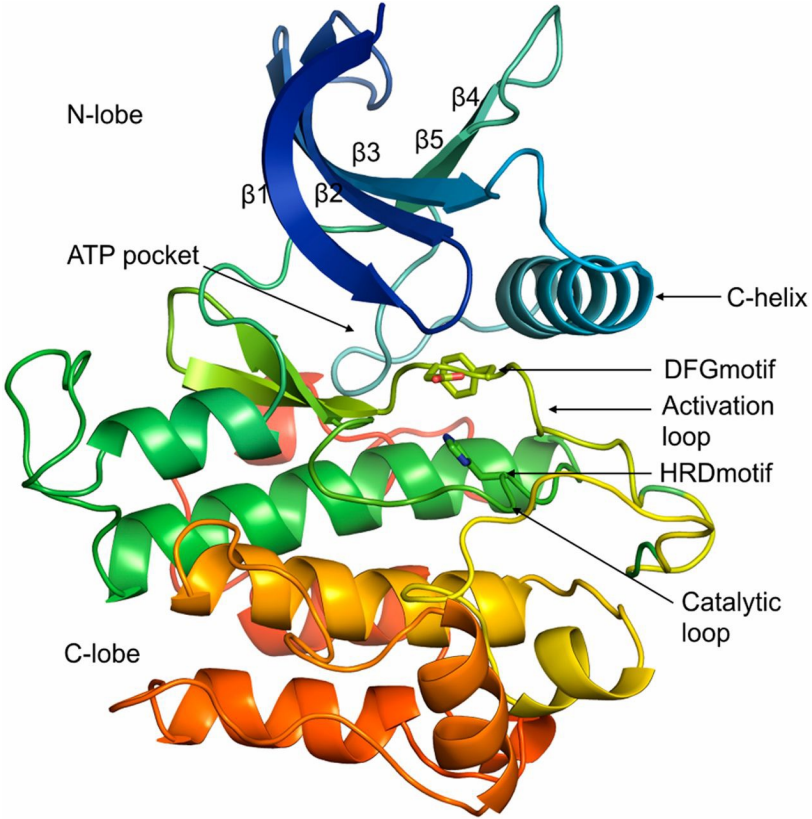
In Humans, there are 494 *typical* kinase domains but only 437 *catalytic* typical kinase domains

437 catalytic typical kinase domains

57 pseudokinase domains

481 genes

494 domains



Insulin receptor kinase

Pseudokinases are protein kinase domains that are missing key elements that facilitate catalysis.

Kinase Structures in the Protein Data Bank

http://dunbrack.fccc.edu/kincore

Active forms of kinases should be able to bind ATP, Mg ions, and *substrate*

Models of active kinases are needed to:

- Understand substrate specificity
- Differences between active and inactive conformations of each kinase
- Understand regulation of catalysis of each kinase
- Druggability of the active form of the kinase
- Effect of mutations on kinase activity

Can we make AlphaFold2 models of all 437 kinases in their active form?

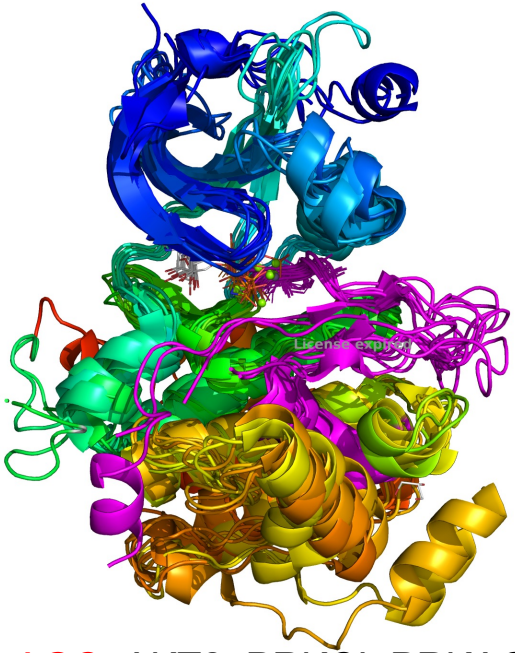
What does an active kinase look like anyway? How many are there in the PDB?

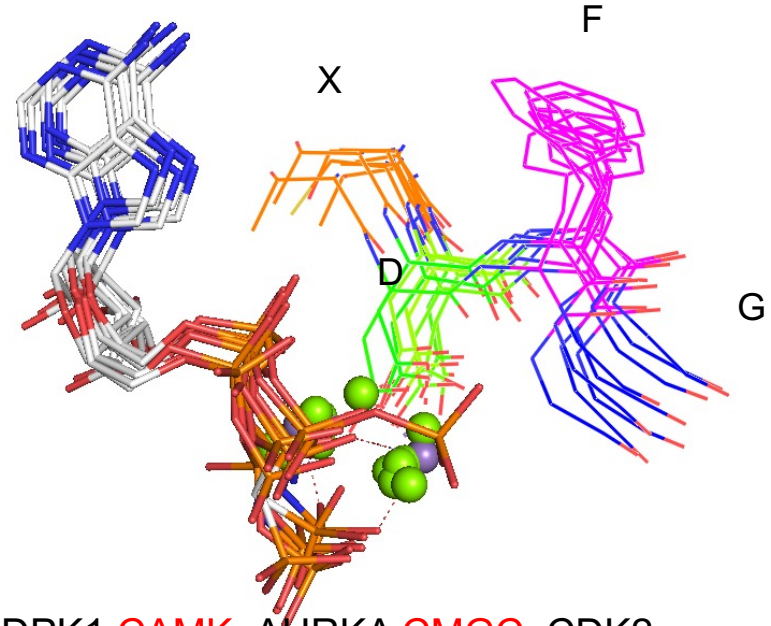
Look at substrate-bound structures and ATP-bound structures

What makes a kinase structure active?

”Catalytically primed structures” (Modi and Dunbrack, PNAS, 2019

- ATP-bound
- Ion complex (Mg^{2+} or Mn^{2+})
- Activation loop phosphorylated
- Resolution ≤ 2.5 Å





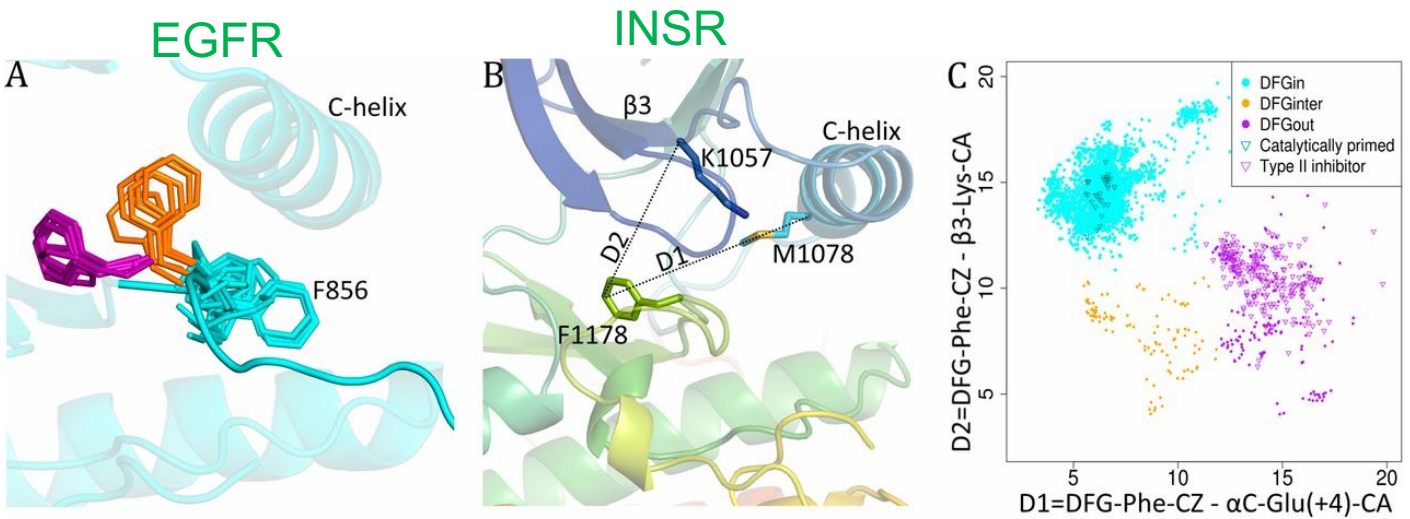
AGC_AKT2, PRKCI, PRKACA, PDPK1 CAMK_AURKA CMGC_CDK2
STE PAK1 PAK4, STK24 TYR FGFR2 INSR

What makes a kinase structure “active”?

Common, well-known features of active kinases observed in these structures:

1. DFGin – Phe near C-helix
2. BLAminus – dihedral angles of XDFG motif (Phe in g- rotamer)
3. Salt bridge in N-terminal domain (C-helix Glu + Beta3 Lys)

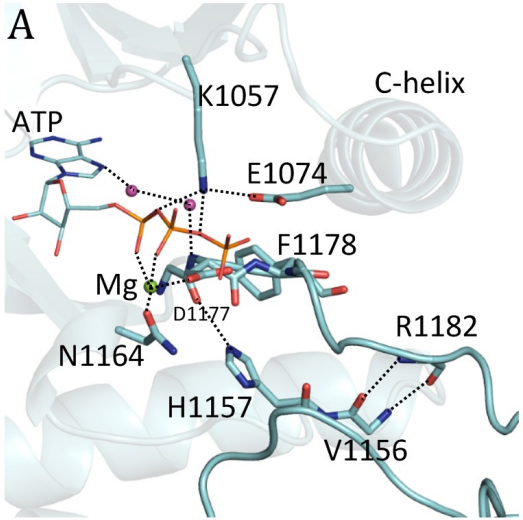
DFGout
DFGinter
DFGin

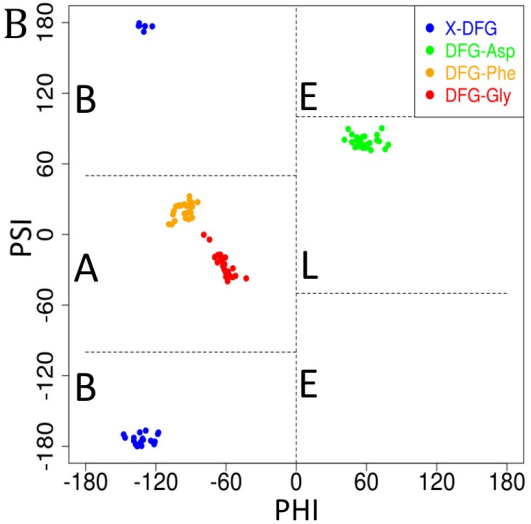


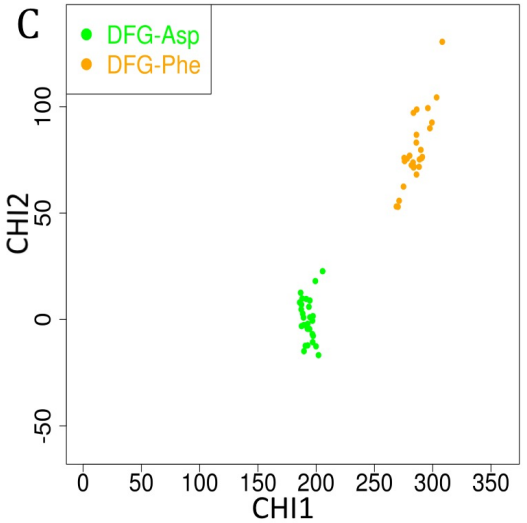
Catalytically Primed Structures Have Uniform Dihedral Angles of X-DFG Motif

Modi and Dunbrack, PNAS 2019

1. DFGin – Phe near C-helix
2. BLAminus – dihedral angles of XDFG motif (Phe in g- rotamer)
3. Salt bridge in N-terminal domain (C-helix Glu + Beta3 Lys)

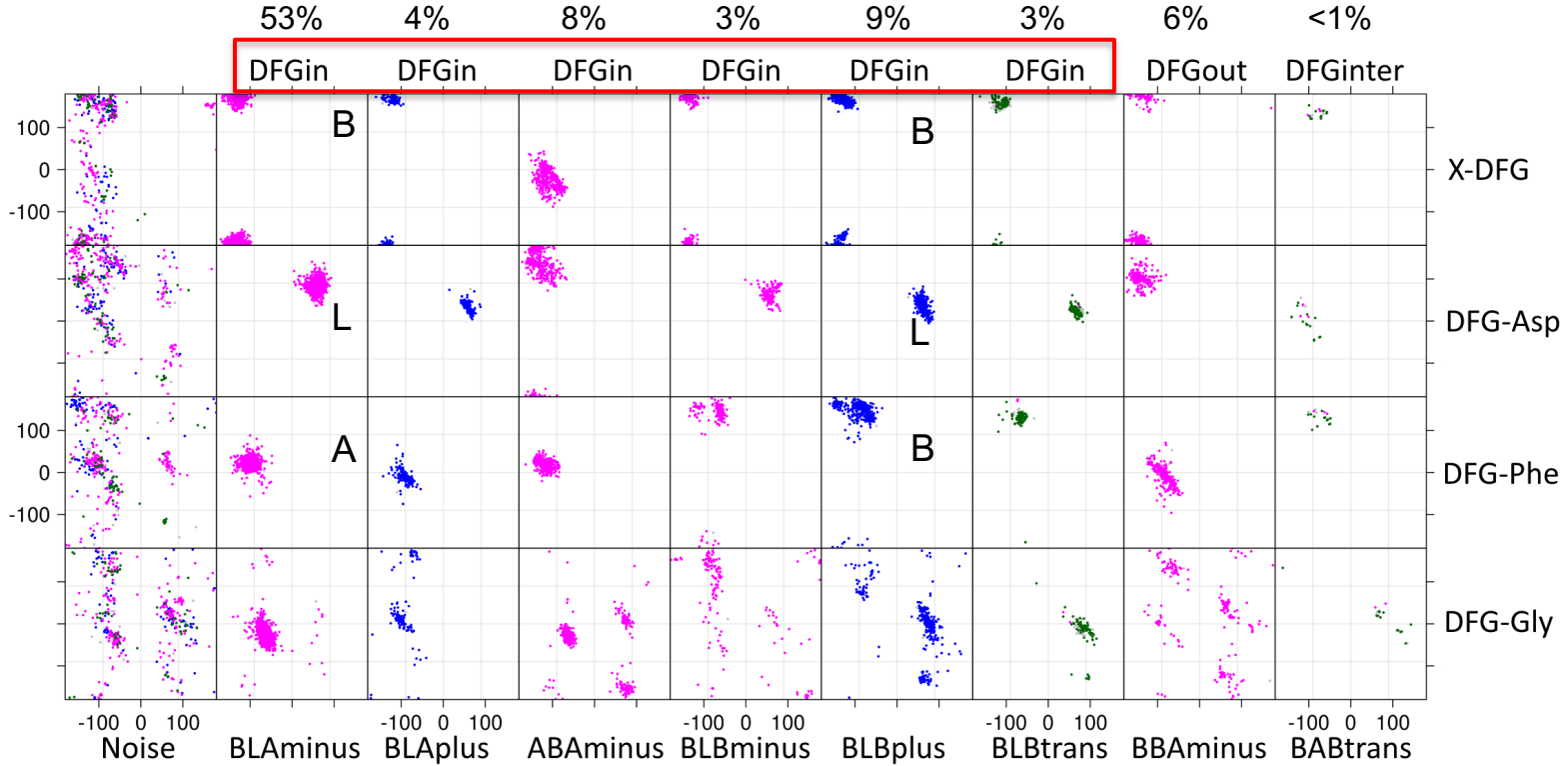







```
X = B conformation
D = L conformation (t rotamer)
F = A conformation (g- rotamer)
G = A conformation
```

Clustering in dihedral angle space of XDF ϕ,ψ and Phe χ_1 with DBSCAN



Nomenclature: Rama of X-D-F residues and rotamer of Phe

BLAminus = active (or really necessary but not sufficient for activity)

BLBplus = "Src-inactive"

BLBtrans = "CDK-inactive"

BBAminus=Type II inhibitor binding

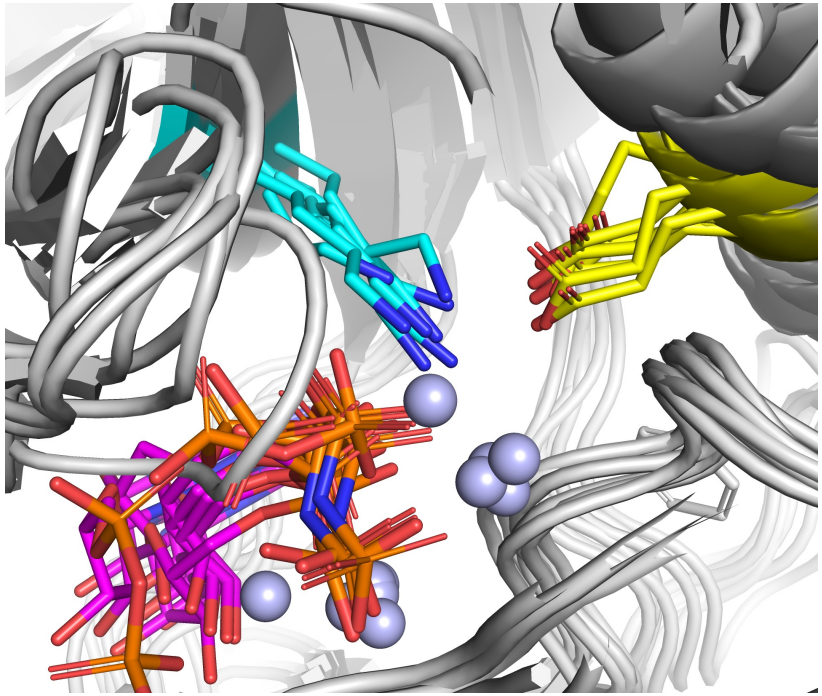
g-minus

g-plus

trans

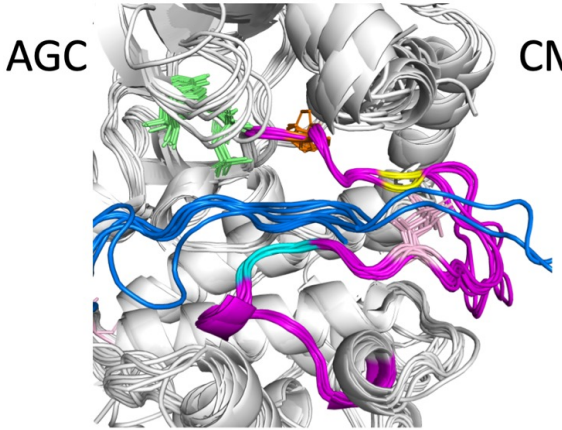
Salt bridge in the N-terminal domain positions ATP

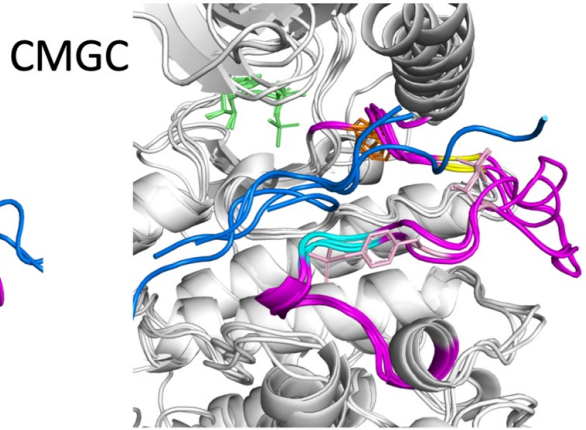
1. DFGin – Phe near C-helix
2. BLAminus – dihedral angles of XDFG motif (Phe in g- rotamer)
3. Salt bridge in N-terminal domain (C-helix Glu + Beta3 Lys) chelates ATP phosphate

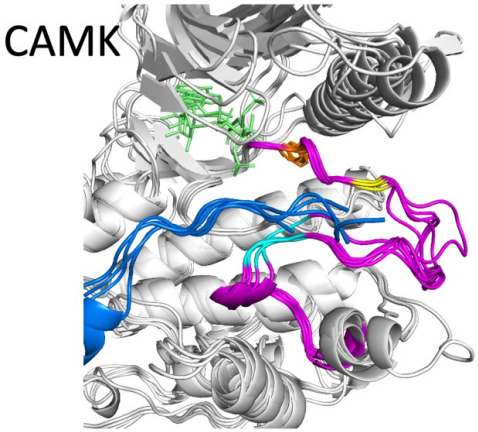


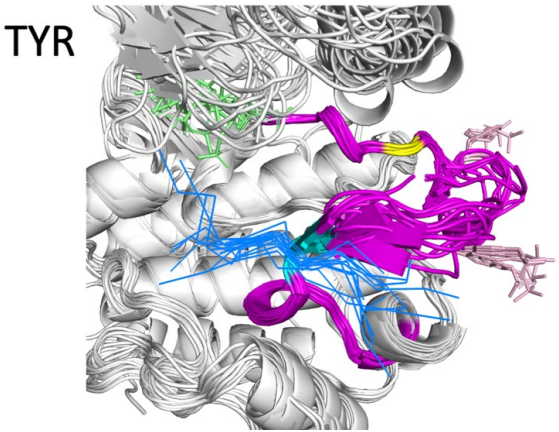
What else makes a kinase structure “active”?

Look at 40 substrate-bound structures (most with ATP and Mg^{2+})









40 Unique kinase/substrate pairs:

	Peptide	Protein
AGC	6	2
CAMK	3	2
CK1	1	
CMGC	3	1
OTHER	2	
STE	1	1
TKL	1	1
TYR	10	6

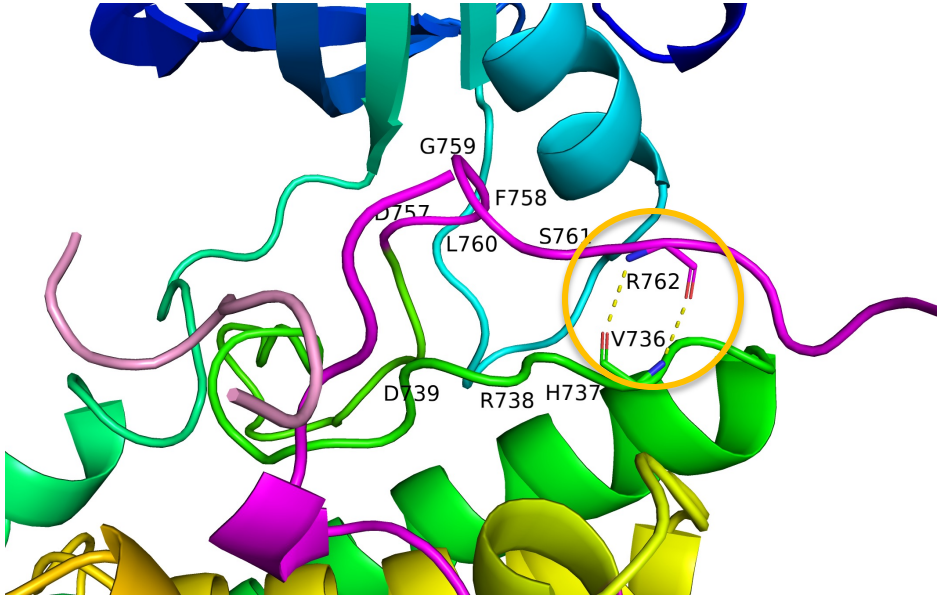
What else makes a kinase active?

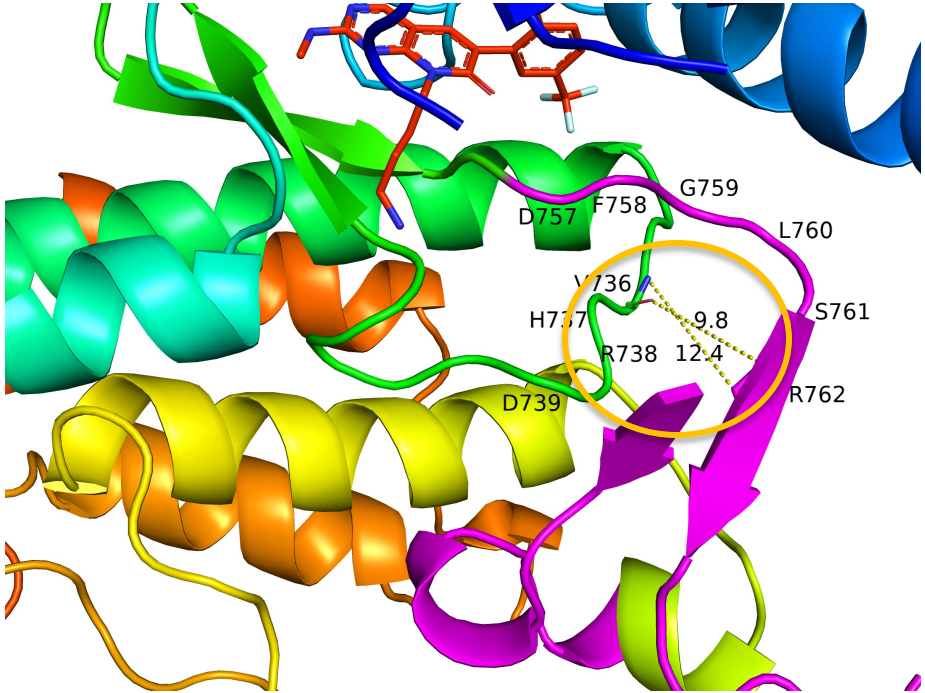
Contacts between substrate and activation loop in red

Kinase	PDB	Ligand	DFGxxx	*xxxxx*CGTxxxxAPE
AGC_AKT1	4ekk	ANP-MG	DFG**4**56	*54**3210987654***APE
AGC_AKT2	1o6l	ANP-MG	DFG**4**56	*543**210987654***APE
AGC_PRKACA	7e0z	ANP-MG	DFG**4**56	*5432**10987654***APE
AGC_PRKCI	5lih	ADP-MG	DF**G45**6	*543**210987**654APE
CAMK_CAMK2A	7uir	ATP-MG	**D**FG**4**56	*5432**10987654***APE
CAMK_PHGK1	2phk	ATP-MG	DF**G4**56	*543**210987654***APE
CAMK_PIM1	2bzk	ANP-MG	DFG**4**56	*5432**10987**654APE
CK1_CSNK1D	6ru7	ADP	**DFG456**	*543**210987**654APE
CMGC_CDK2	1qmz	ATP-MG	**D**FG**45****6**	*543**210987**654APE
CMGC_DYRK1A	2wo6	CAZ	DFG**4**56	*543**210987654***APE
OTHER_CDC7	6ya7	ADP-ZN	**12345****6**	*543**210987**654APE
STE_PAK1	4jdi	ANP-MG	DFG**4**56	*543**210987654***APE
TYR_ABL1	2g2i	ADP	-	*543**210987**654APE
TYR_EGFR	5czh	-	**D**FG456	*543**210987**654APE
TYR_EPHA2	3fxx	ANP-MG	123**4**56	*5432**10987**654APE
TYR_FES	3cbl	STU	DFG**4**56	*543**210987**654APE
TYR_FGFR2	2pvf	ACP-MG	**D**FG**4**56	*5432**10987**654APE
TYR_IGF1R	1k3a	ACP	DFG**4**56	***543210987654***APE
TYR_INSR	3bu5	ATP-MG	**D**FG**4**56	***543210987654***APE
TYR_SYK	5c27	50J	**D**FG**4**56	*543**210987**654APE
Consensus			DFG1-6	APE4-13

XHRD – DFG6 backbone hydrogen bond distance

TYR kinase

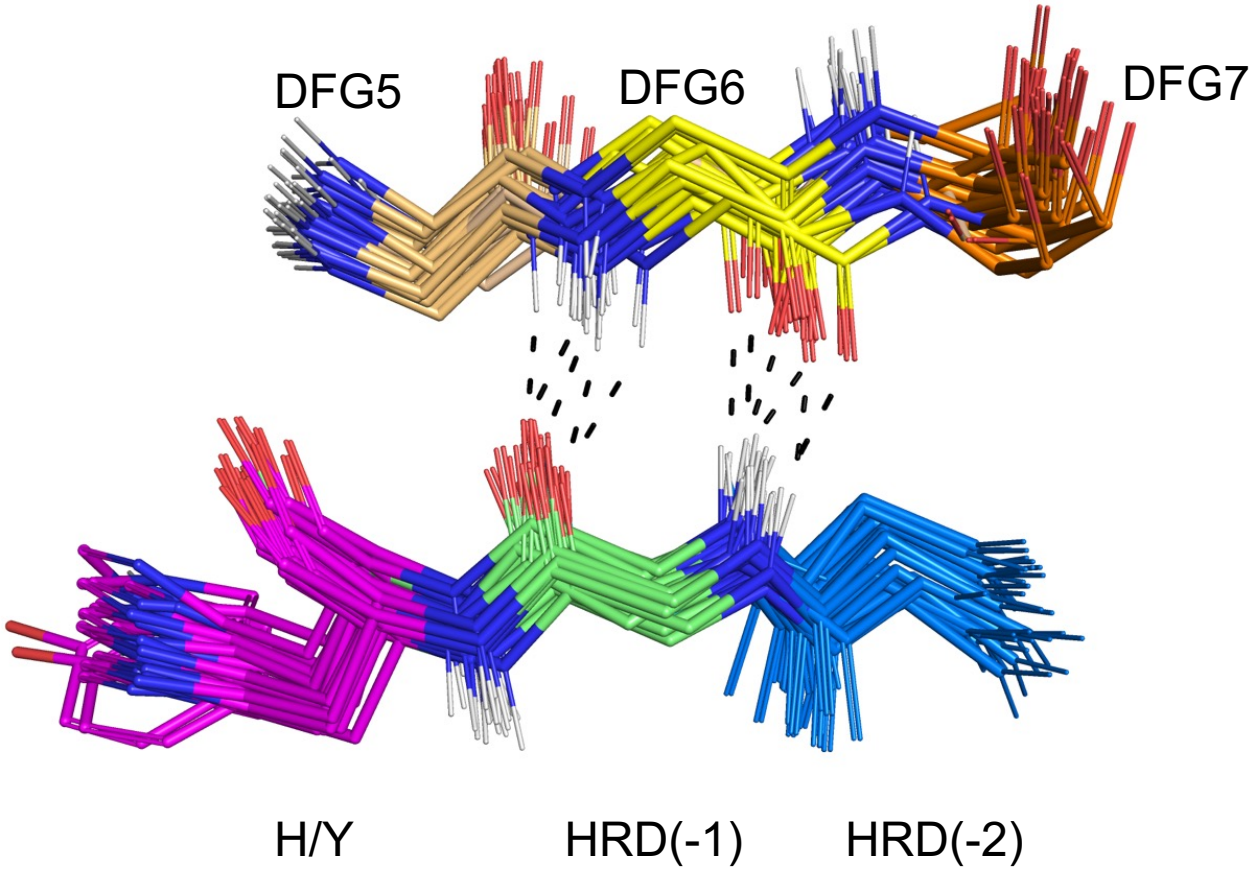




EPHA2 3fy2 with substrate
BLAminus ActLoopNT-in (3.0 Å)

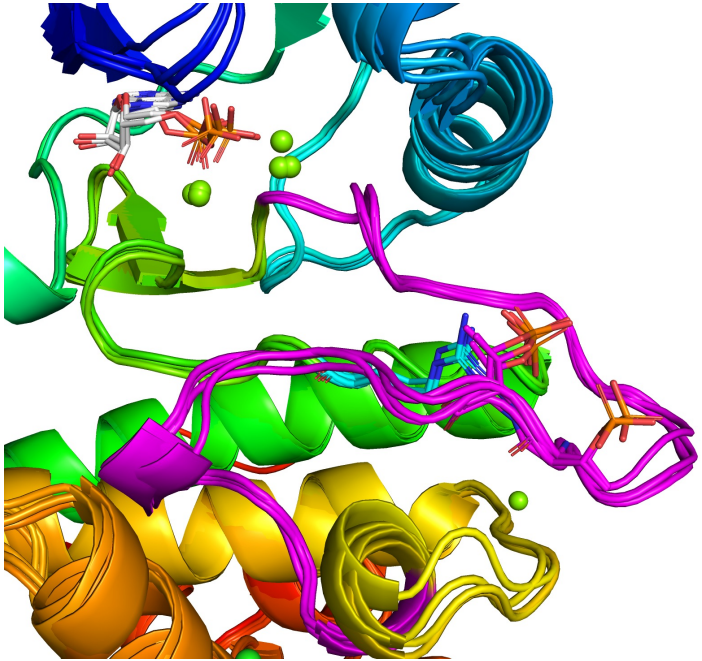
EPHA2 8bio, no substrate
BLAminus ActLoopNT-out (9.8Å)

DFG6-Xhrd backbone-backbone hydrogen bonds in substrate-bound structures

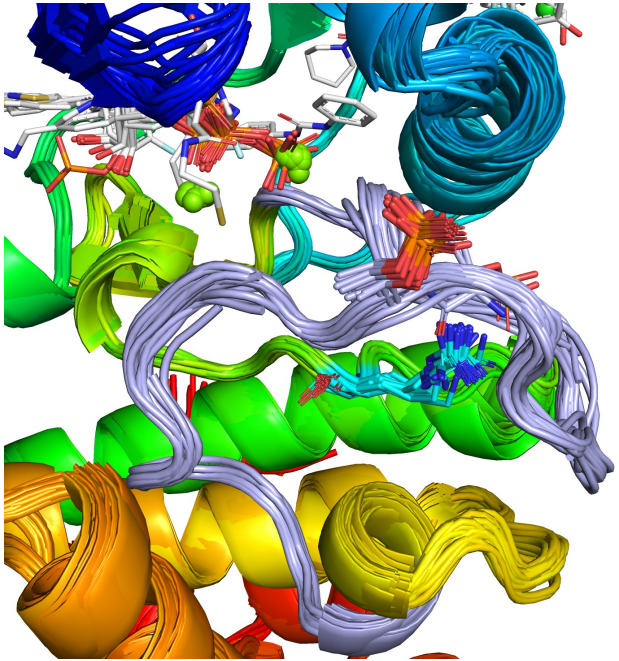


C-terminal Activation Loop Conformation

AURKA BLAminus Structures



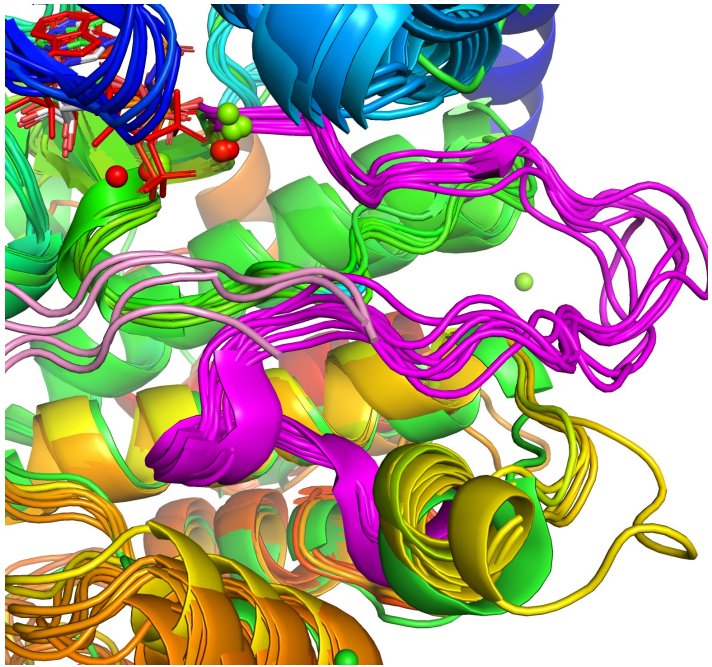
AURKA Conf1



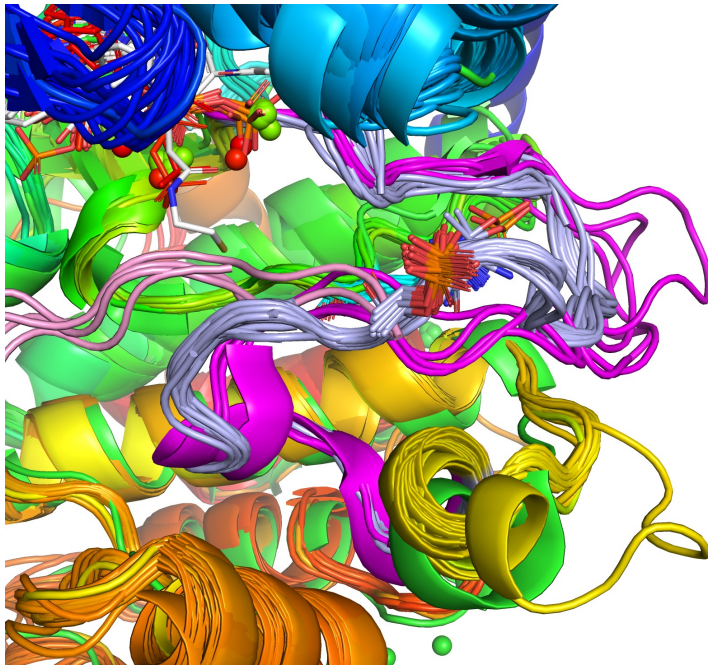
AURKA Conf2

Conf1 is substrate-binding – Conf2 is not

AURKA BLAminus Structures

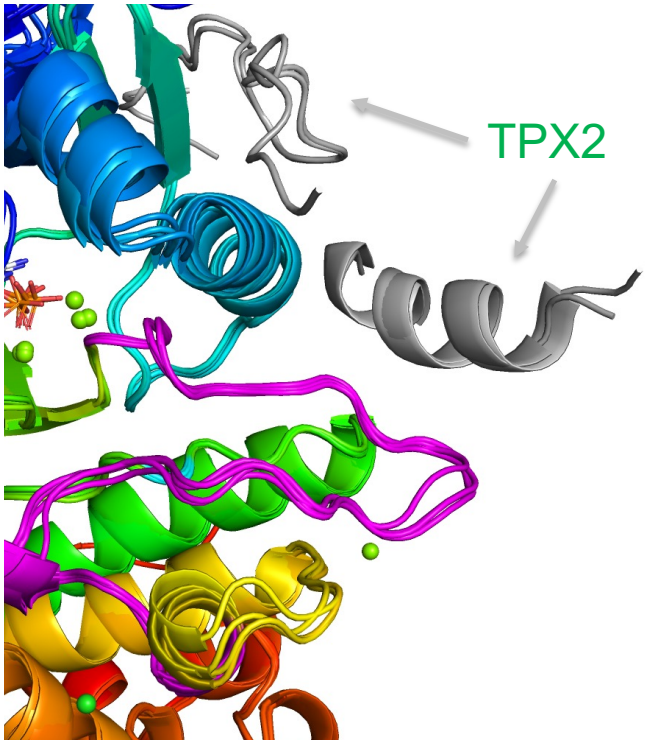


AURKA Conf1 (TPX2-bound)
with substrate-bound AGC_PRKACA,
CAMK_CAMK2A, PIM1, PHGK1

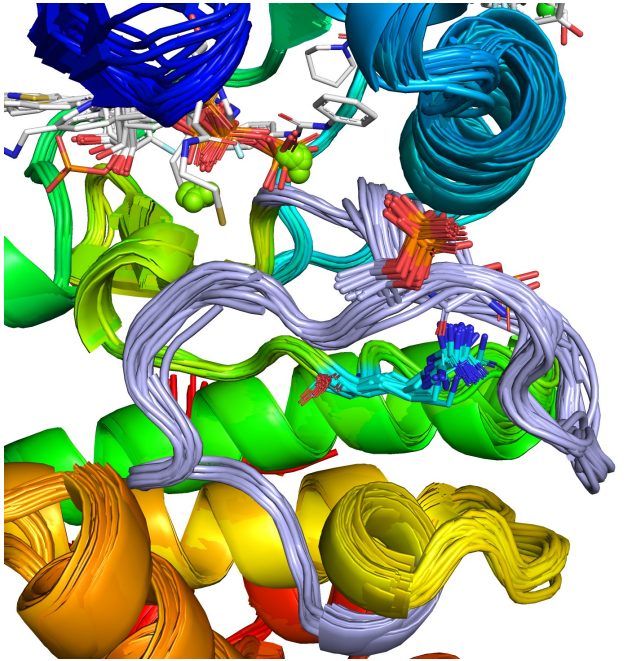


AURKA Conf2 with
substrate bound
structures **(clash)**

TPX2 makes AURKA active by “pulling on the activation loop”



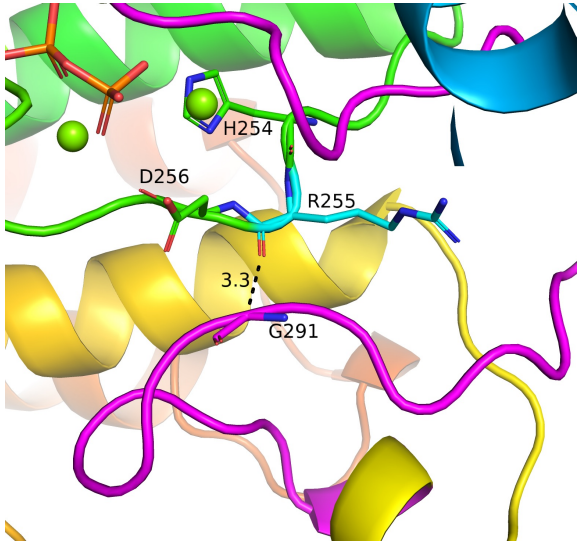
AURKA Conf1



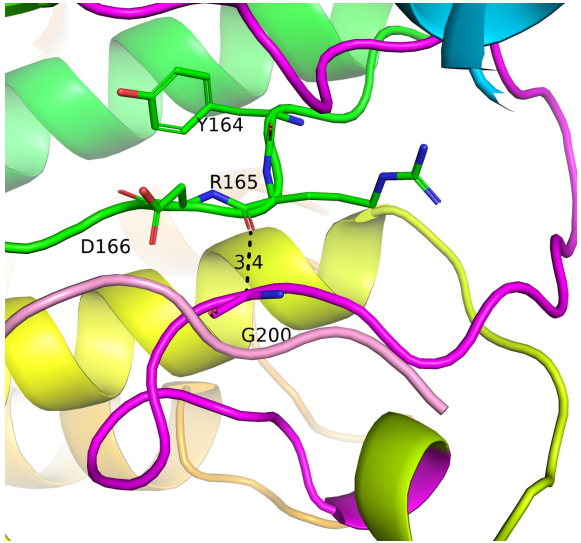
AURKA Conf2

Criterion for the C-terminus of the Activation Loop

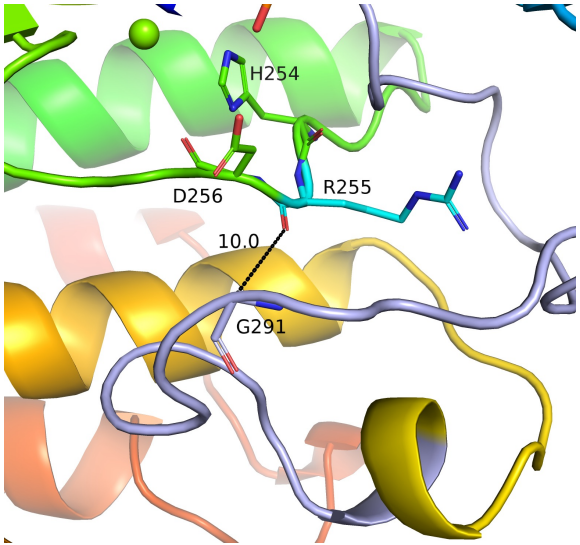
APE9(Cα)-hRd(O) distance criterion ≤ 6.0 Å



AURKA Conf1 distance
5LXMA, 3.3 Å
cGtldylPPE

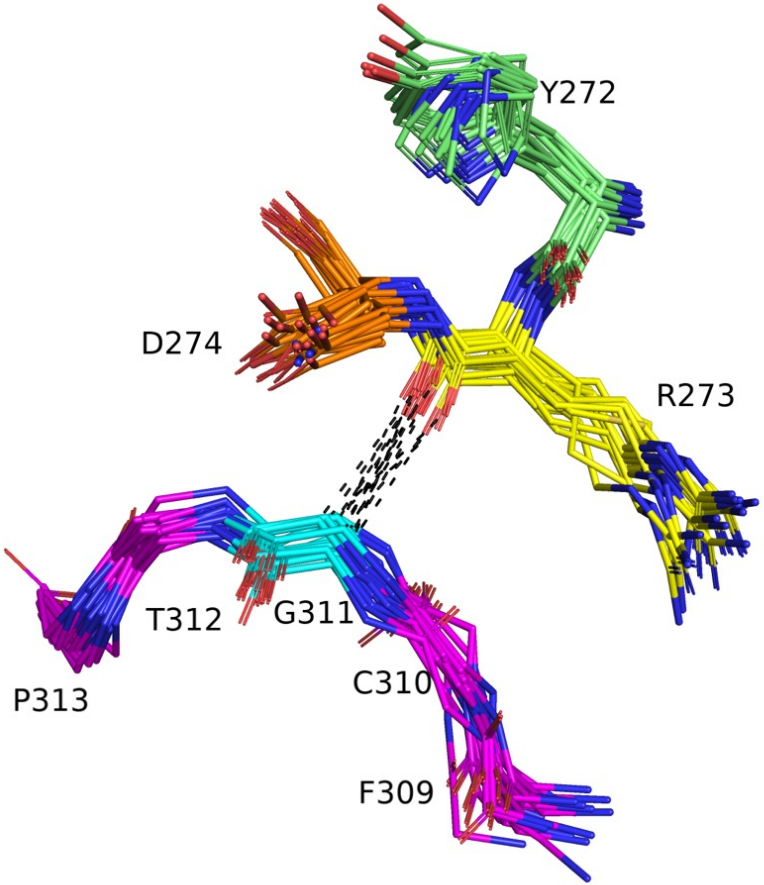


AGC_PRKACA substrate bound
structure (7E0Z), dist=3.4 Å
cGtpeylAPE



AURKA Conf2 distance
5DR2A, 10.0 Å
(substrate clash)

APE9(Cα)-hRd(O) distance in substrate-bound structures



5 Active Criteria Applied to Catalytic Kinases PDB

ActLoopCT (APE9 distances) has biggest effect after DFGin-BLAminus

1. DFGin
2. BLAminus
3. SaltBr-in: distance<3.6Å
4. ActLoopNT-in: DFG6-Xhrd distance<3.6
5. ActLoopCT-in: APE9/hRd distance<6.0 (nonTYR); <8.0 TYR)

All catalytic human kinases	**437**	**100%**
Any conformational state	268	61%
DFGin+BLAminus	202	46%
DFGin+BLAminus+SaltBr-in	188	43%
DFGin+BLAminus+ActLoopNT-in	193	44%
DFGin+BLAminus+ActLoopCT-in	**162**	**37%**
Active kinase structures	**155**	**36%**
Active structures with full Actloop	**130**	**30%**

Making models of active structures of all 437 human catalytic protein kinases with AlphaFold2 with shallow sequence alignments and active templates

Sequence sources for MSA

1. Orthologues of query (>50% seqid, >90% coverage, <90% identity to each other
2. Sequences in the same kinase family (AGC, CAMK, etc.)
3. Uniprot90

Number of sequences in MSA:

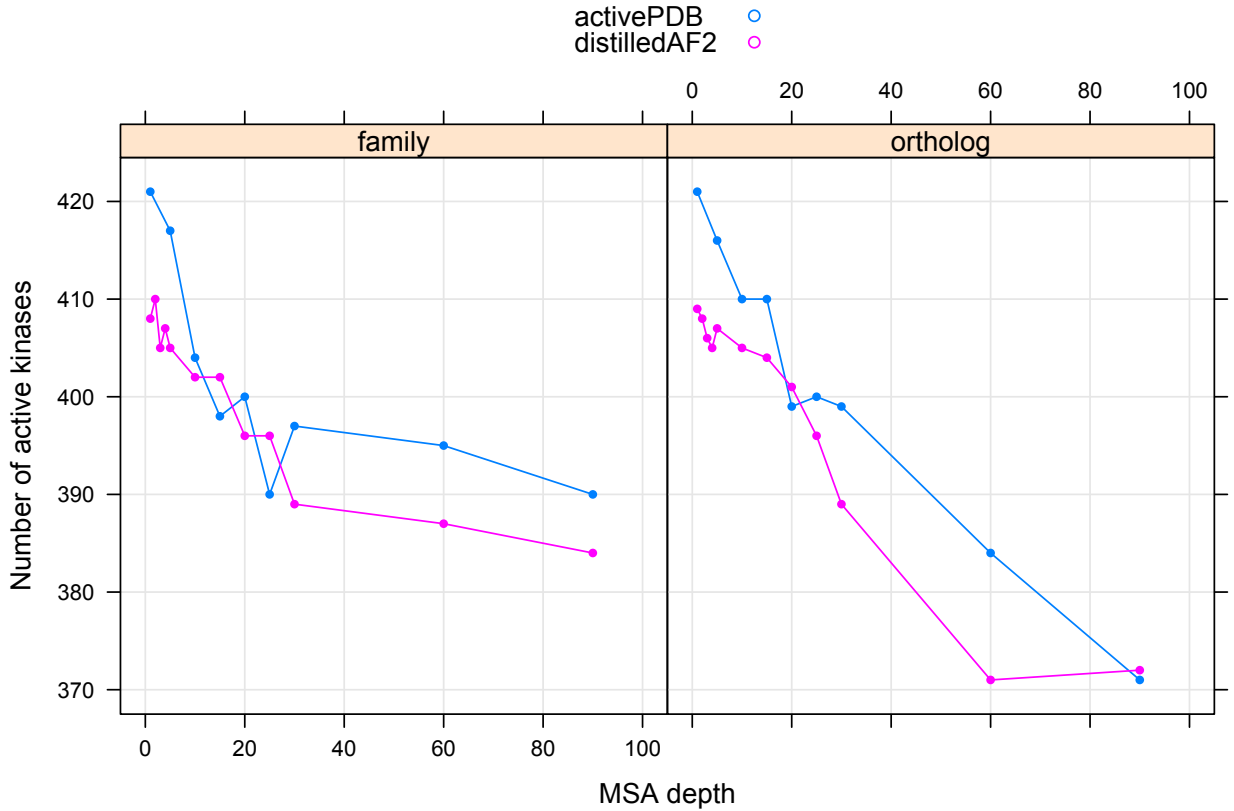
5, 10, 15, 20, 25, 30, 60, 90,10000 (for Uniprot90)

Template databases [skip template for modeling same protein]

1. Active structures in PDB (according to 5 criteria) = 165 kinases and 278 chains (human and nonhuman)
2. “Distillation templates”: 246 active models built by AlphaFold2

Heo and Feig, *Proteins* 2022; Del Alamo et al, *eLife, 2022*

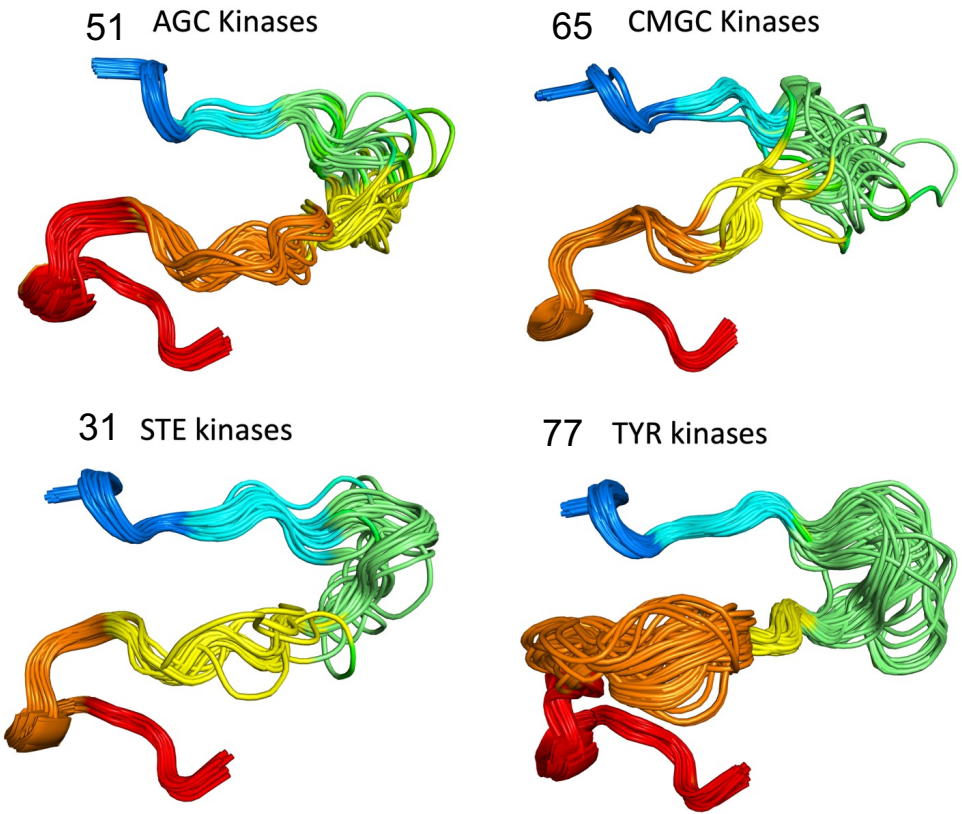
AlphaFold2 Active Models of 437 Human Catalytic Kinases



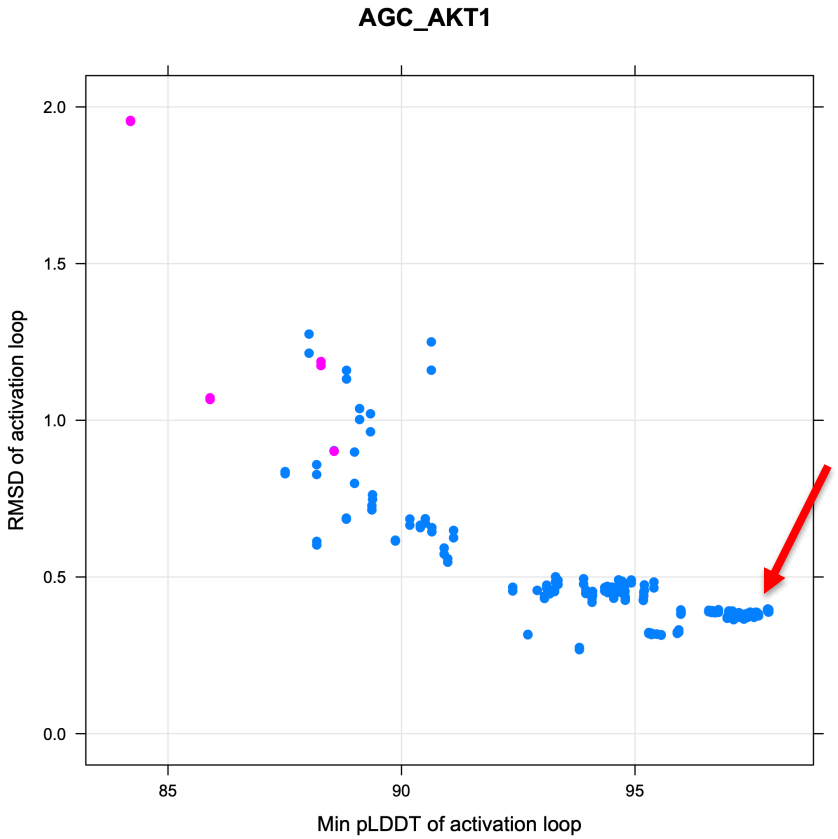
EBI models:
only 209/437

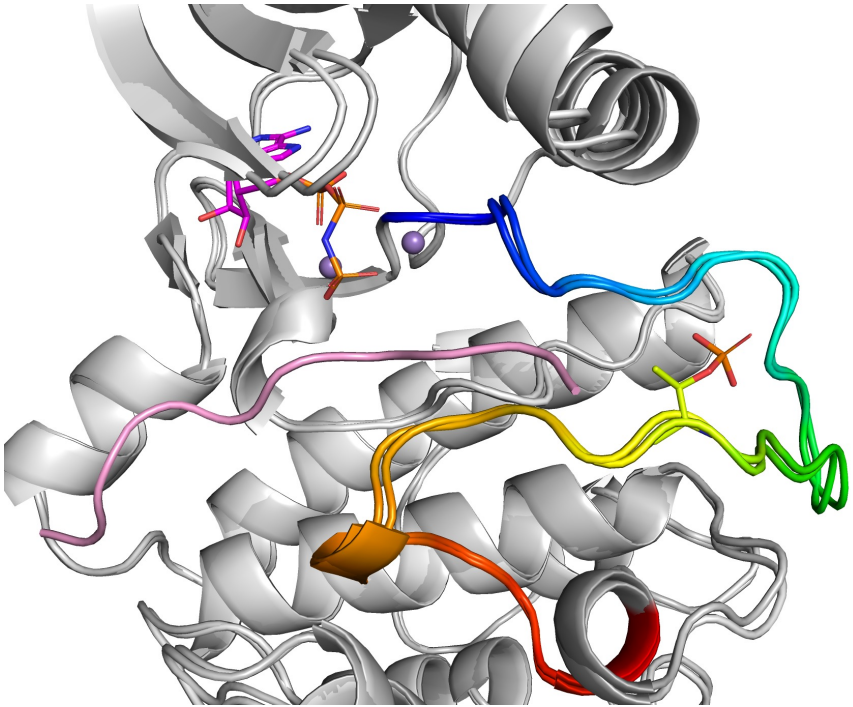
Problem children: TYR_LMTK2, CAMK_OBSCN-2

Combining all MSA sources and depths with PDB and AF2 templates (and one mutant) → 437 active kinase models

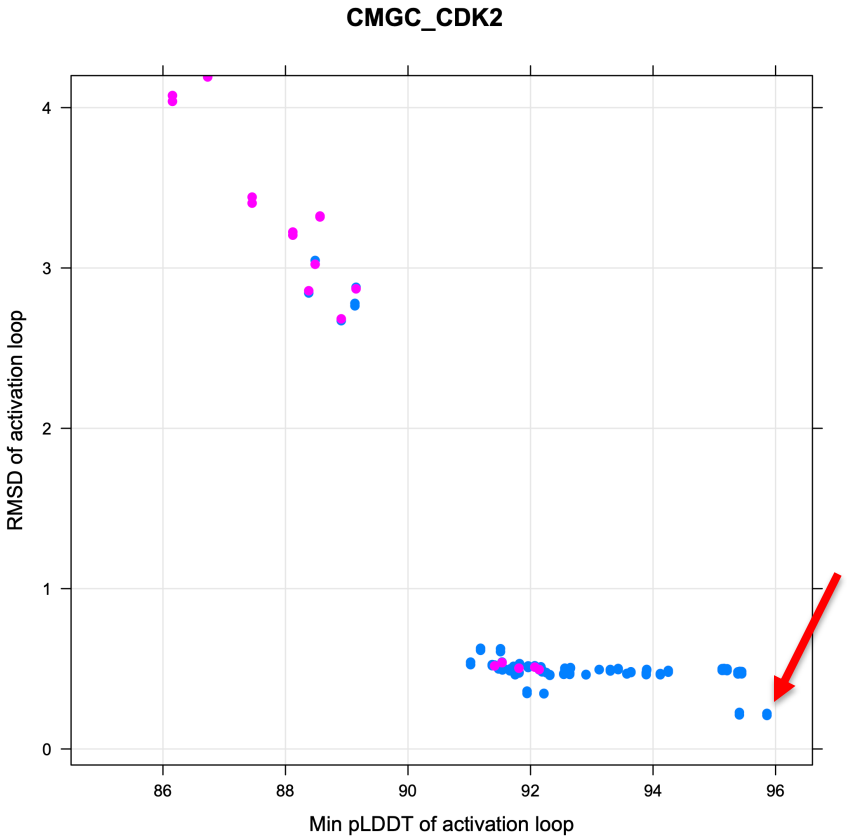


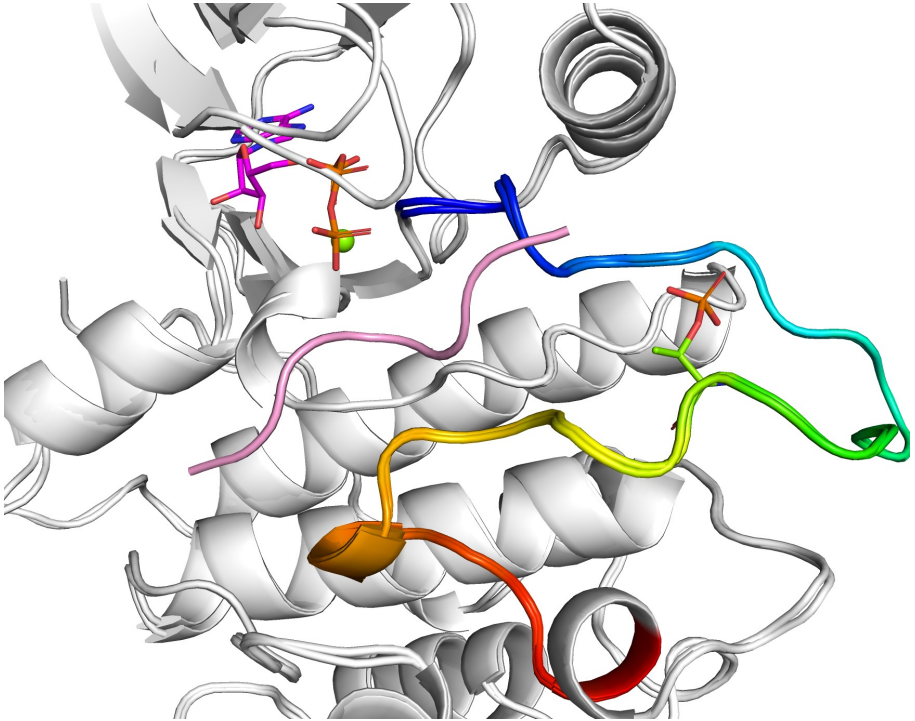
Picking the best active model with min pLDDT of activation loop





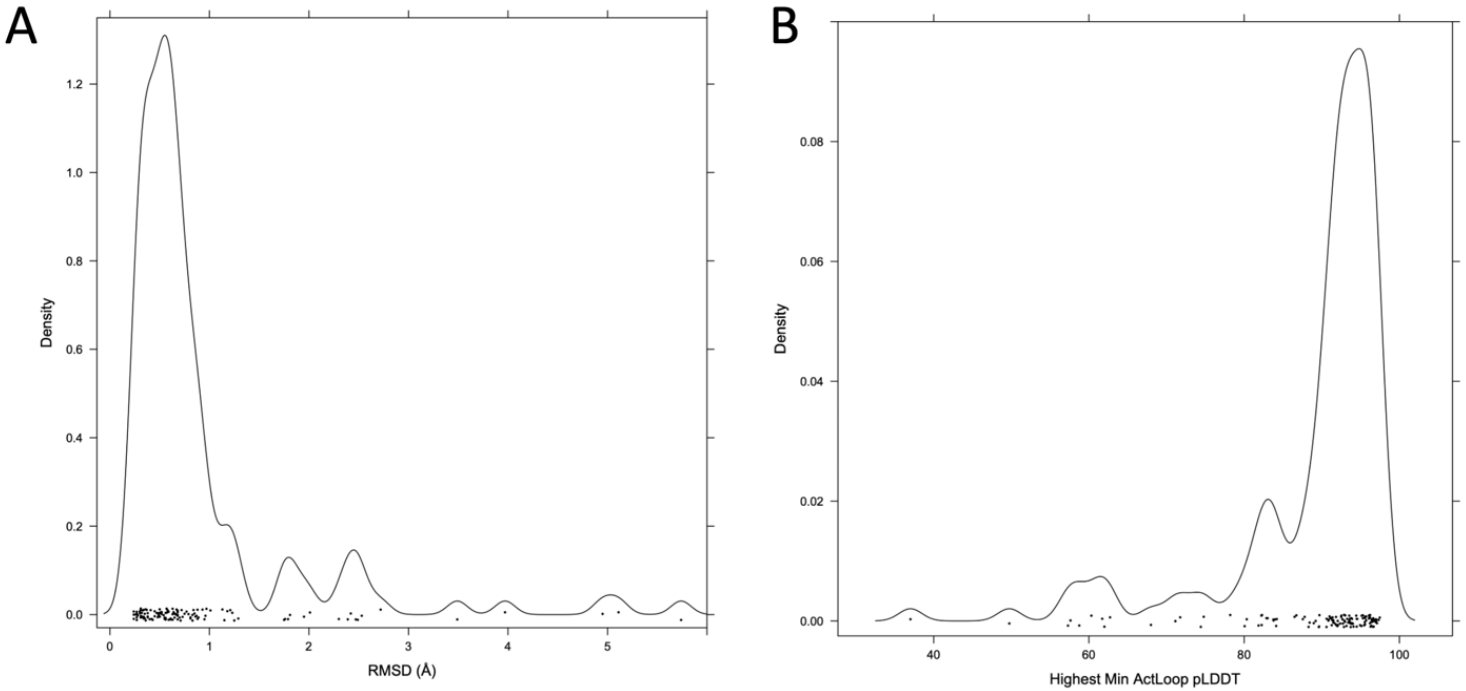
Picking the best active model with min pLDDT of activation loop





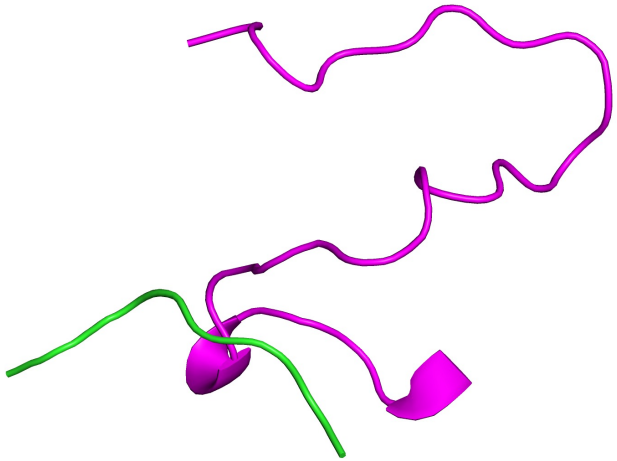
Benchmark of 130 “Active” Structures from PDB with complete activation loops

80% better than 1.0 Å, 90% better than 2.0Å

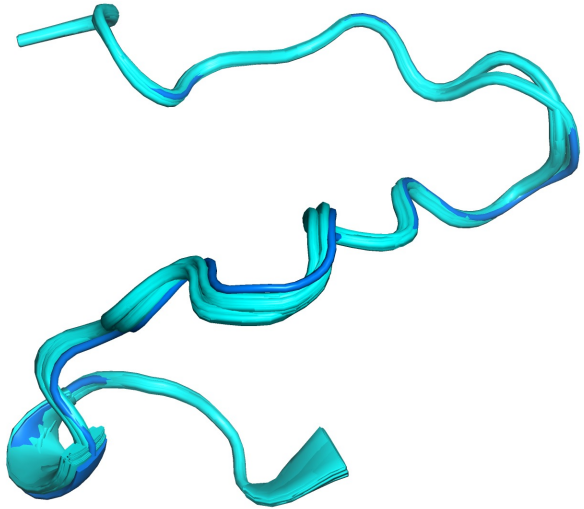


SRC in the benchmark

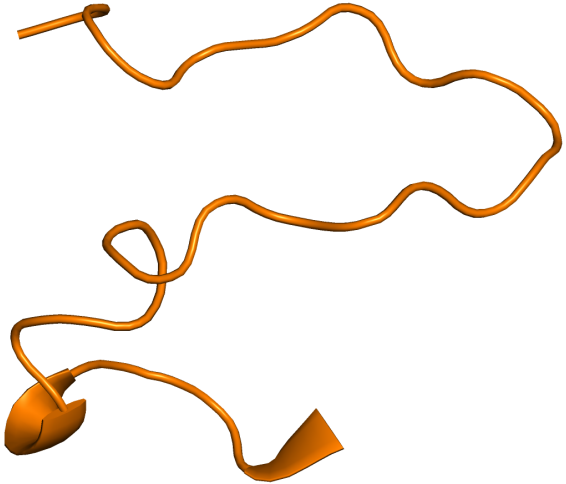
No human active SRC structure in PDB but SRC_CHICK is active in PDB 3DQW



ABL1
substrate (2G2I)



SRC_CHICK (3DQW)
SRC_HUMAN (AF2)



SRC_HUMAN (1Y57)
Used in MD simulations
of “Active” SRC
RMSD = 2.5 Å to AF2

Summary and Prospects

- Rigorous structural bioinformatics defines features of substrate-bound kinase structures (“Active”)
  - DFGin
  - BLAminus
  - Saltbridge
  - ActLoop-NT position
  - ActLoop-CT position

- Only one third of 437 kinases are active in PDB

- We produced AF2 models of all 437 in active form with shallow sequence alignments, active templates, heavy sampling, and strict structural bioinformatics criteria

- Inactive states are more complicated and more varied: most kinases will not exist in all inactive forms (SRCinactive-BLBplus, DFGout-BBAminus, etc.)

- AlphaFold-Multimer is capable of binding substrates to kinases and modeling other PPI that regulate kinases. **Our models can be used as templates.**

- http://dunbrack.fccc.edu/kincore/activemodels for data on PDB and (eventually) AF2 models

Acknowledgments

Bulat Faezov
Vivek Modi (now GSK)

NIH MIRA R35 GM122517

http://dunbrack.fccc.edu/kincore/activemodels

Twitter: @RolandDunbrack

- Structural bioinformatics
- Rants about people parking in bike lanes.
- LGBTQ issues in science