Defining a new nomenclature for the structures of active and inactive kinases

Vivek Modi[a] and Roland L. Dunbrack Jr.[a,1]

[a]Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, PA 19111

Edited by Kevan M. Shokat, University of California, San Francisco, CA, and approved February 15, 2019 (received for review August 20, 2018)

Targeting protein kinases is an important strategy for intervention in cancer. Inhibitors are directed at the active conformation or a variety of inactive conformations. While attempts have been made to classify these conformations, a structurally rigorous catalog of states has not been achieved. The kinase activation loop is crucial for catalysis and begins with the conserved DFGmotif. This motif is observed in two major classes of conformations, DFGin—a set of active and inactive conformations where the Phe residue is in contact with the C-helix of the N-terminal lobe—and DFGout—an inactive form where Phe occupies the ATP site exposing the C-helix pocket. We have developed a clustering of kinase conformations based on the location of the Phe side chain (DFGin, DFGout, and DFGinter or intermediate) and the backbone dihedral angles of the sequence X-D-F, where X is the residue before the DFGmotif, and the DFG-Phe side-chain rotamer, utilizing a density-based clustering algorithm. We have identified eight distinct conformations and labeled them based on the Ramachandran regions (A, alpha; B, beta; L, left) of the XDF motif and the Phe rotamer (minus, plus, trans). Our clustering divides the DFGin group into six clusters including BLAminus, which contains active structures, and two common inactive forms, BLBplus and ABAminus. DFGout structures are predominantly in the BBAminus conformation, which is essentially required for binding type II inhibitors. The inactive conformations have specific features that make them unable to bind ATP, magnesium, and/or substrates. Our structurally intuitive nomenclature will aid in understanding the conformational dynamics of kinases and structure-based development of kinase drugs.

protein kinases | structural bioinformatics | cell signaling

Phosphorylation is a fundamental mechanism by which signaling pathways are regulated in cells. Protein kinases are cellular sentinels which catalyze the phosphorylation reaction by transferring the γ-phosphate of an ATP molecule to Ser, Thr, or Tyr residues of the substrate. Due to their crucial role in the functioning of the cell, protein kinases are tightly regulated. Dysregulation of kinases may result in variety of disorders including cancer, making development of compounds for modulating kinase activity an important therapeutic strategy.

The human genome contains ∼500 protein kinases that share a common fold consisting of two lobes: an N-terminal lobe, consisting of a five-stranded β-sheet with an α-helix called the C-helix, and a C-terminal lobe comprising six α-helices (Fig. 1). They are divided broadly into nine families based on their sequences (1). The two lobes are connected by a flexible hinge region forming the ATP-binding site in the middle of the protein. The active site comprises several structural elements that are crucial for enzymatic activity. The activation loop is typically 20 to 30 residues in length beginning with a conserved DFG motif (usually Asp-Phe-Gly) and extending up to an APE motif (usually Ala-Pro-Glu). In active kinase structures, this loop forms a cleft that binds substrate. Bound substrate peptide forms specific interactions with the conserved HRD motif (usually His-Arg-Asp) which occurs in the catalytic loop of the protein. In the active conformation, the DFG motif Asp is in a position and orientation to bind a magnesium ion that interacts directly with an oxygen atom of the β phosphate of ATP. The active state exhibits an inward disposition of the C-helix which positions a conserved Glu in the helix to form a salt bridge with a Lys residue in the β3 strand. When the salt bridge is formed, the lysine side chain forms hydrogen bonds with oxygen atoms of the α and β phosphates of ATP. The N-lobe has a GxGxxG motif in a loop that stabilizes the phosphates of the bound ATP molecule during catalysis. The catalytically active state of a kinase requires a unique assembly of these elements that create an environment conducive to the phosphotransfer reaction. The regulation of the activity of a kinase is achieved in part by the plasticity of these elements of the structure (2).

Inactive states of a kinase do not have the chemical constraints required for catalytic activity and therefore kinases exhibit multiple inactive conformations (3). Typically, in an inactive conformation the activation loop is collapsed onto the surface of the protein, blocking substrate binding and rendering the kinase catalytically inactive. In addition, many inactive conformations have positions of the DFGmotif incompatible with binding ATP and magnesium ion required for catalysis. In the DFGout conformation, DFG-Phe and DFG-Asp swap positions so that DFG-Phe occupies the ATP binding pocket and DFG-Asp is out of the active site. There are diverse DFGin structures from multiple kinases where DFG-Phe remains adjacent to the C-helix but in a different orientation (and sometimes position) from that of active DFGin structures. There are also structures where the Phe is in positions intermediate between the typical DFGin and DFGout states. The many inactive, non-DFGout conformations have been variously referred to as pseudo DFGout, DFGup, SRC-like inactive, and atypical DFGout (4, 5). Although DFGin and DFGout are broadly recognized groups of conformations, a consensus nomenclature for the inactive states is lacking.

The DFGin and DFGout conformations have been used as the basis of grouping the inhibitors developed against the active site

Significance

Protein kinases play important roles in signaling pathways and are widely studied as drug targets. Their active site exhibits remarkable structural variation as observed in the large number of available crystal structures. We have developed a clustering scheme and nomenclature to categorize and label all the observed conformations in human protein kinases. This has enabled us to clearly define the geometry of the active state and to distinguish closely related inactive states which were previously not characterized. Our classification of kinase conformations will help in better understanding the conformational dynamics of these proteins and the development of inhibitors against them.

Author contributions: R.L.D. designed research; V.M. performed research; V.M. analyzed data; and V.M. and R.L.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data and scripts for this paper are available at Zenodo (doi.org/10.5281/zenodo.2580462).

[1]To whom correspondence should be addressed. Email: roland.dunbrack@fccc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814279116/-/DCSupplemental.

Published online March 13, 2019.

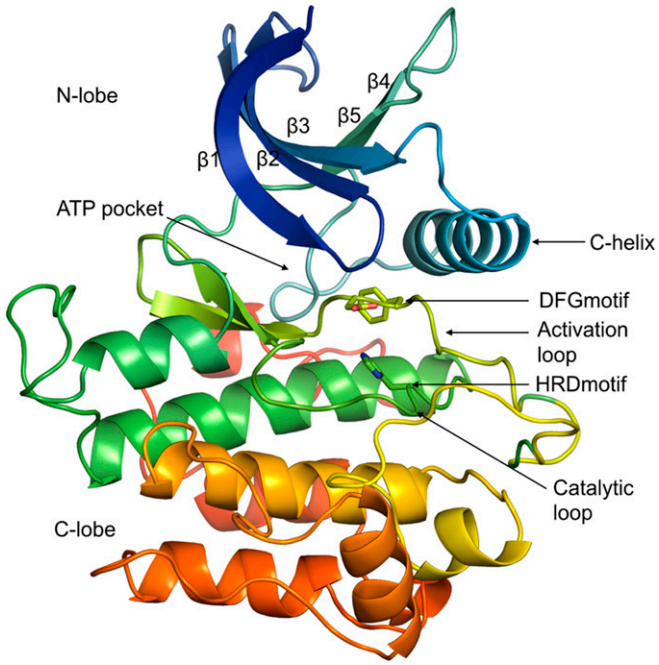

Fig. 1. Structure of a typical protein kinase domain displaying ATP binding site and conserved elements around it (INSR kinase, PDB ID code 1GAG).

of these proteins into two main categories (6, 7). Molecules such as dasatinib which occupy the ATP pocket only are called type I inhibitors and typically bind DFGin conformations, but not exclusively. Type II Inhibitors like imatinib bind to the DFGout state and extend into the hydrophobic allosteric pocket underneath the C-helix (8). Design of better inhibitors could be guided by a better understanding and classification of the conformational variation observed in kinases.

There have been some attempts to classify kinase structures in the Protein Data Bank (PDB) (now over 3,300) and to study inhibitor interactions (9–12). Möbitz (11) has performed a quantitative classification of all of the mammalian kinases using pseudo dihedral angles of four consecutive Cα atoms of the residues of the DFGmotif and its neighbors and its distance from the C-helix. This resulted in a scheme dividing kinase conformations into 12 categories with labels "FG-down," "FG-down αC-out," "G-down αC-out," "A-under P BRAF," "A-under P-IGF1R," and so on. Recently, Ung et al. (12) used a similar idea of using two directional vectors for the DFGmotif residues and the distance from the C-helix to classify kinases into five groups, C-helix-in-DFGin (CIDI), C-helix-in-DFGout (CIDO), C-helix-out-DFGin (CODI), C-helix-out-DFGout (CODO), and ωCD. Some other classification schemes have emphasized the binding modes of inhibitors (4, 13).

In this paper, we present a clustering and classification of the conformational states of protein kinases that addresses some of the deficiencies of previous such efforts. These deficiencies include failing to distinguish DFGin inactive conformations from active structures, either too few or too many structural categories, and an inability to automatically classify new structures added to the PDB. In the current work, we have clustered all of the human kinase structures at two levels of structural detail. First, at a broader level we grouped kinase structures into three categories depending on the spatial position of the DFG-Phe side chain. These three groups are labeled the DFGin, DFGout, and DFGinter (intermediate) conformations. Second, we clustered each of the three spatial groups at a finer level based on the dihedral angles required to place the Phe side chain: the backbone dihedral angles ϕ and ψ of the residue preceding the DFGmotif (X-DFG), the DFG-Asp residue, and the DFG-Phe residue, as well as the χ_1 side-chain dihedral angle of the DFG-Phe residue. This produced a total of eight clusters—six for DFGin and one cluster each for the DFGout and DFGinter groups.

We have developed a nomenclature that is intuitive to structural biologists based on the regions of the Ramachandran map occupied by the X, D, and F residues of the X-DFG motif ("A" for alpha-helical region, "B" for beta-sheet region, and "L" for left-handed helical region) and the χ_1 rotamer of the Phe side chain ("minus" for the −60° rotamer, "plus" for the +60° rotamer, and "trans" for the 180° rotamer). We have clearly identified the active state of kinases, designated "BLAminus,"' which is the most common kinase conformation in the PDB. Further, we also clearly define different inactive DFGin conformations which were previously grouped together. The most common inactive DFGin conformations are BLBplus and ABAminus. The type II-binding DFGout state is labeled BBAminus. Overall, our clustering and nomenclature scheme provides a structural catalog of human kinase conformations which will provide deeper insight into the structural variation of these proteins, benefitting structure-guided drug design.

Results

Clustering Kinase Conformations Based on Spatial Location of the DFG-Phe Residue. Human protein kinases in the PDB were identified by sequence, excluding proteins such as PI3–PI4 kinases that are distantly related to canonical protein kinases but possess highly divergent folds. This led to a dataset with 244 human kinase domain sequences with known structures from 3,343 PDB entries, having 4,834 polypeptide chains containing kinase domains (some asymmetric units have multiple copies of the kinase). To identify a set of reliable structures where the catalytic machinery is primed to catalyze the phosphorylation reaction and the activation loop orientation is conducive to substrate binding we selected structures that satisfy the following criteria: (*i*) resolution ≤2.25 Å and R-factor ≤0.25; (*ii*) ATP or a triphosphate analog of ATP bound to the active site (PDB ligand codes ATP, ANP, ACP, or AGS); (*iii*) Mg^{2+} or Mn^{2+} ion bound in the active site; and (*iv*) a phosphorylated Ser, Thr, or Tyr residue in the activation loop. Not all kinases require a phosphorylated residue in the activation loop for activity, but its presence is usually indicative of an active kinase structure. This led to the identification of a set of 28 chains from 24 PDB entries and 12 different kinases (listed in *SI Appendix*, Table S1). We refer to these structures as "catalytically primed." An example is shown in Fig. 2*A*. Dihedral angle features of these structures are plotted in Fig. 2 *B* and *C* and discussed further below.

In protein kinases, the active structure and various inactive structures are distinguished by the wide variety of positions and conformations of the residues of the DFG motif. The well-known DFGin and DFGout classes describe the rough position of the Asp and Phe residues of the DFG motif but fail to capture how these positions are attained. Intermediate states between DFGin and DFGout have been described (14). These three groups of structures are easily distinguished in many kinases as shown for EGFR in Fig. 3*A*.

To capture the location of the DFG-Phe residue, we calculated its distance from two conserved residues in the N-terminal domain (Fig. 3*B*): D1 = dist(αC-Glu(+4)-Cα, DFG-Phe-Cζ) and D2 = dist(β3-Lys-Cα, DFG-Phe-Cζ). D1 is the distance between the Cα atom of the fourth residue after the conserved Glu residue in the C-helix (Exxx**X**) and the outermost atom of the DFG-Phe ring (Cζ). Even as the αC-helix moves outward, the hydrophobic Exxx**X** residue does not move significantly since it is located closer to the pivot point of the helix toward the back of the kinase N-terminal domain. The distance of this residue to the Phe ring serves to distinguish DFGin structures, where the Phe ring is adjacent to or under the C-helix, from DFGout structures, where the ring has moved a substantial distance laterally from the C-helix. D2 is the distance between the Cα atom of the conserved Lys residue from the β3 strand to the Cζ atom of the DFG-Phe side chain. It captures the closeness of DFG-Phe to the N-lobe β-sheet strands, thus giving an estimate of the upward

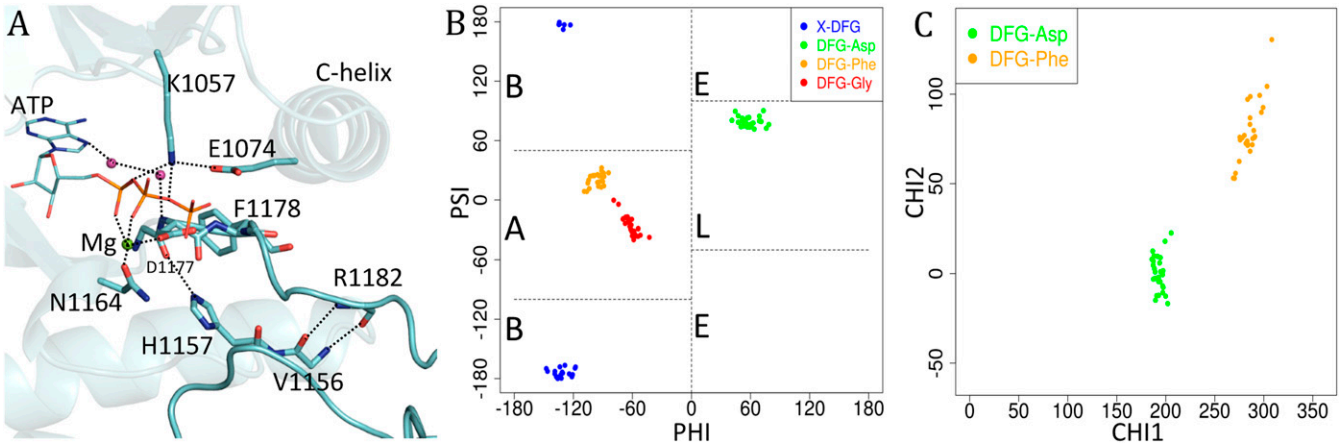

Fig. 2. Analysis of catalytically primed structures. (*A*) Catalytically primed structure of insulin receptor kinase (PDB ID code 3BU5) with bound ATP and Mg^{2+} ion. Important interactions are marked with dashed lines. (*B*) Ramachandran plot of catalytically primed structures (those with bound ATP and Mg^{2+}, a phosphorylated activation loop, and resolution ≤2.25 Å): X-DFG (residue before the DFGmotif, blue), DFG-Asp (green), DFG-Phe (orange), and DFG-Gly (red). The Ramachandran regions are marked A (alpha), B (beta), L (left), and E (epsilon). (*C*) Scatterplot of side-chain dihedral angles χ_1 and χ_2 for DFG-Asp (green) and DFG-Phe (orange) of catalytically primed structures. Asp is in a trans rotamer ($\chi_1 \sim 180°$) and Phe is in a g^- rotamer ($\chi_1 \sim 300°$ or $-60°$).

position of the Phe ring, distinguishing DFGin conformations from structures where the Phe ring is in an intermediate position between DFGin and DFGout.

These distances are plotted against each other in Fig. 3*C*. We have clustered these distances into three groups using average linkage hierarchical clustering. The choice of three groups in clustering algorithm was guided by the visual inspection of large number of structures suggesting three broad regions or pockets occupied by DFG-Phe residue (Fig. 3*A*). Based on this we have classified the kinase structures into the following three groups:

i) DFGin: This is the largest group, consisting of 4,333 chains (89.6%) from 227 kinases shown in cyan-colored points in Fig. 3*C*, representing the DFG motif orientations where DFG-Phe is packed against or under the C-helix (cyan in Fig. 3*A*). It consists of many related conformations with the typical DFGin active orientation forming the largest subset of this group. All of the catalytically primed structures belong to this group (cyan triangles in Fig. 3*C*).

ii) DFGout: This is the second-largest group, consisting of 388 chains (8%) from 60 kinases, displayed in purple-colored points representing the structures where DFG-Phe is moved into the ATP binding pocket (purple in Fig. 3*A*). The structures with a type II inhibitor bound form a subset of this group (purple triangles in Fig. 3*C*).

iii) DFGinter (DFGintermediate): This is the smallest group, consisting of 113 chains (2.3%) from 27 kinases, in which the DFG-Phe side chain is out of the C-helix pocket but has not moved completely to a DFGout conformation (orange in Fig. 3*A* and orange dots in Fig. 3*C*). In most of these cases DFG-Phe is pointing upward toward the β-sheets dividing the active site into two halves. Dodson et al. (14) had previously referred to this state in Aurora A kinase as "DFGup." Recently, Ung et al. (12) have also identified this set of conformations and labeled them as "ωCD."

Clustering Kinase Conformations Based on the Backbone of Activation Loop. For the DFG-Phe side chain to exhibit such wide-ranging localization within the kinase domain fold, the backbone and side-chain dihedral angles leading up to the Phe side chain must be divergent. By examining a large number of structures, we observed

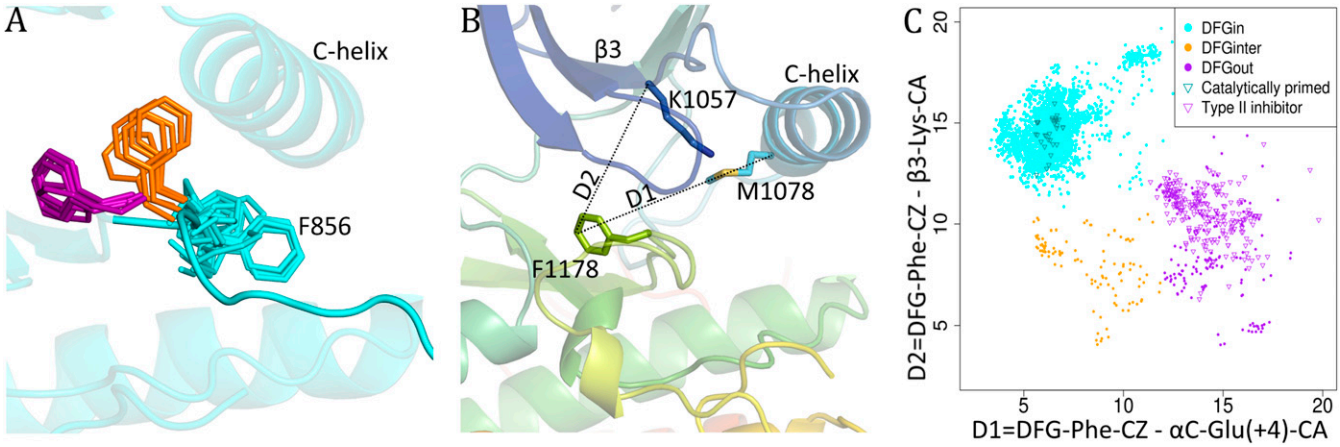

Fig. 3. Locations of Phe side chain in DFGin, DFGinter, and DFGout structures of kinases. (*A*) Positions of DFG-Phe side chain in DFGin (cyan), DFGinter (orange), and DFGout (purple) of EGFR. (*B*) Distances, D1 and D2, used to identify the location of the Phe side chain of the DFGmotif. (*C*) Scatterplot of D1 and D2. DFGin (cyan), DFGinter (orange), and DFGout (purple) structures identified with hierarchical clustering; catalytically primed structures (cyan triangles) and structures with bound type II inhibitors (purple triangles) are marked.

that the structural variation of the activation loop begins with the residue that precedes the DFG motif ("X-DFG"). The position of the Phe side chain is then determined by the backbone dihedrals φ and ψ of X-DFG, DFG-Asp, and DFG-Phe as well as the first side-chain dihedral of the DFG-Phe residue.

The high-resolution catalytically primed structures have very precise values of these dihedral angles (Fig. 2 *B* and *C*). The backbone conformations of the X-D-F residues of catalytically primed structures occupy the beta (B), left (L), and alpha (A) regions of the Ramachandran map, respectively (as defined in Fig. 2*B*). The DFG-Phe side chain adopts a χ_1 gauche-minus rotamer ($\chi_1 \sim -60°$ or equivalently 300°; $\chi_2 \sim 90°$, orange dots in the χ_1-χ_2 scatterplot in Fig. 2*C*) and points slightly downward into a pocket underneath the C-helix (Fig. 2*A*). In all of these structures, the DFG-Asp adopts a χ_1 trans rotamer (~180°) and the χ_2 dihedral is ~0° (green dots in Fig. 2*C*). This places the Asp carboxylate atoms in a horizontal orientation to chelate an Mg^{2+} /Mn^{2+} ion on one side and to form a hydrogen bond with the NH of DFG-Gly on the other. The Mg^{2+} ion forms a tight interaction with an oxygen atom on the β-phosphate group (Fig. 2*A*).

Based on these observations on the catalytically primed structures, we decided to cluster the conformations of kinase structures using a metric based on the backbone dihedrals φ and ψ of X-DFG, DFG-Asp, and DFG-Phe as well as the first side-chain dihedral of the DFG-Phe residue. The backbone of the DFG-Gly residue exhibits high flexibility and therefore was not included in the clustering. Each kinase chain is represented by a vector of these seven dihedrals. The distance between these vectors is calculated by a metric from directional statistics (15), which we used previously in our work on clustering antibody CDR loop conformations (16). The distance matrix calculated with this metric was used as input to DBSCAN (density-based spatial clustering of applications with noise), which is a density-based clustering algorithm (17). DBSCAN groups together data points that are connected by high density with each other while identifying the points in low-density regions as noise. We clustered the DFGin, DFGout, and DFGinter groups of structures separately.

For the DFGin group we obtained six clusters with DBSCAN (Fig. 4). The Ramachandran map dihedral angles and side-chain dihedral angles naturally cluster in high-density regions, and DBSCAN readily identified conformations of the X-DFG motif residues that conform to well-known populations (18). By using the Ramachandran region annotation (A, B, L, and E) for the X, D, and F residues and the DFG-Phe χ_1 rotamer (minus = −60°, plus = +60°, and trans = 180°), these clusters are labeled as BLAminus, BLAplus, ABAminus, BLBminus, BLBplus, and BLBtrans. Example structures are shown in Fig. 5 *A*–*F*. All of the catalytically primed structures are observed in the BLAminus cluster; the Gly conformation for these residues is uniformly in the alpha (A) position (Fig. 2*B*). Although the DFG-Gly residue dihedrals were not used in the clustering, for most of the clusters there is one dominant conformation of this residue.

For the DFGout group we obtained just one cluster. In this cluster, the X-D-F residues occupy the B-B-A regions of the Ramachandran map (Figs. 4 and 5*G*) and DFG-Phe is in a −60° rotamer, while the Gly residue occupies all four Ramachandran conformations (A, B, L, and E). The cluster is therefore labeled BBAminus. Eighty-two percent of 244 type II inhibitor-bound chains in 187 PDB entries of 42 different kinases are observed in this cluster; the remainder are in the DFGout noise group.

The structures in the DFGinter conformation display more variability than the other states. For the DFGinter group we obtained only one cluster of 20 chains from eight kinases. The X-D-F residues are in a B-A-B conformation (Figs. 4 and 5*H*) and the DFG-Phe residue is observed in a trans rotamer with a few chains displaying a rotamer orientation between g-minus and trans (six chains of CDK2 with DFG-Phe $\chi_1 \sim -100°$). The Gly residue is in an L conformation. Owing to the more prominent side-chain orientation, we have labeled this cluster BABtrans.

We assigned 110 noise points to backbone clusters if the distance between these points and the nearest cluster centroid was less than a certain cutoff (*Methods*). The resulting clusters can be validated by their silhouette scores (*SI Appendix*, Fig. S1). This leaves us with a total of 447 chains (9%) which could not be assigned to any backbone clusters. Although these structures do not get a backbone cluster label, they still belong to a specific spatial group: DFGin (48% of 447 chains), DFGout (31%), or DFGinter (21%). Cluster assignments and associated data for all human kinase chains in the PDB (except those missing the X-DFG

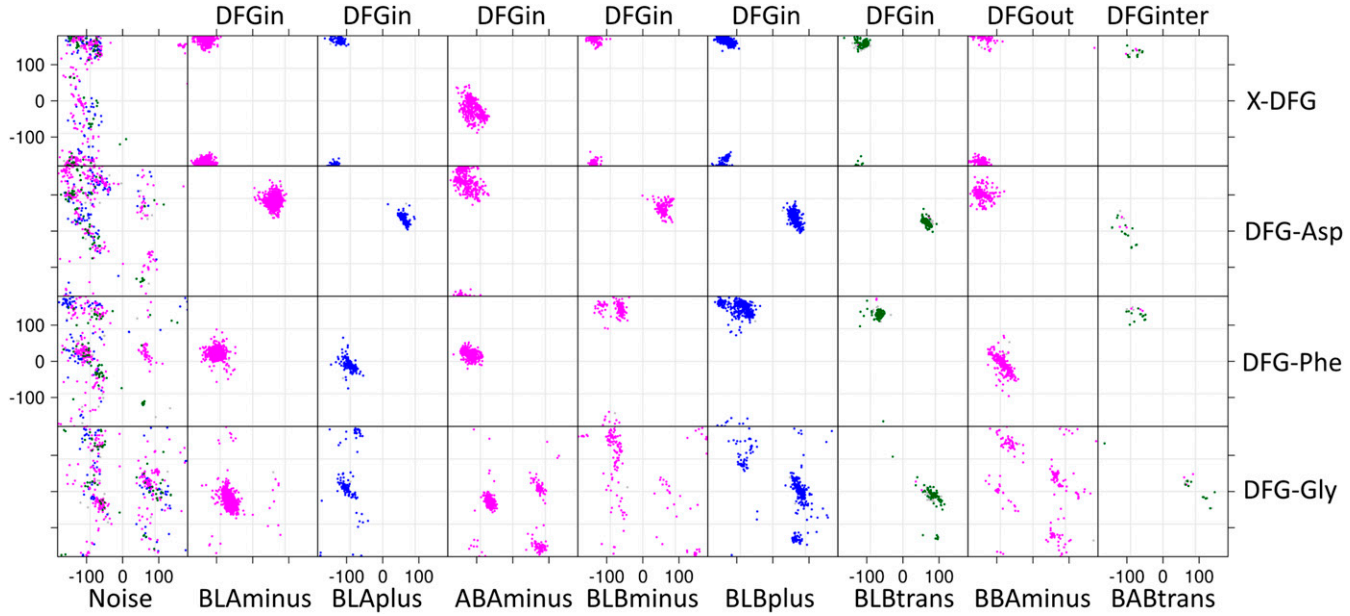

Fig. 4. DBSCAN clustering of dihedral angles that position the DFG-Phe side chain. Clustering was performed on the DFGin, DFGinter, and DFGout groups separately with an angular metric of the backbone dihedral angles of the X-DFG, DFG-Asp, and DFG-Phe residues and the χ_1 of DFG-Phe. Clusters are named by the Ramachandran regions of the X-DFG, DFG-Asp, and DFG-Phe residues and the χ_1 rotamer of DFG-Phe: g^- ($\chi_1 \sim -60°$, minus, magenta); g^+ ($\chi_1 \sim +60°$, plus, blue); trans rotamer ($\chi_1 \sim 180°$, trans, green).

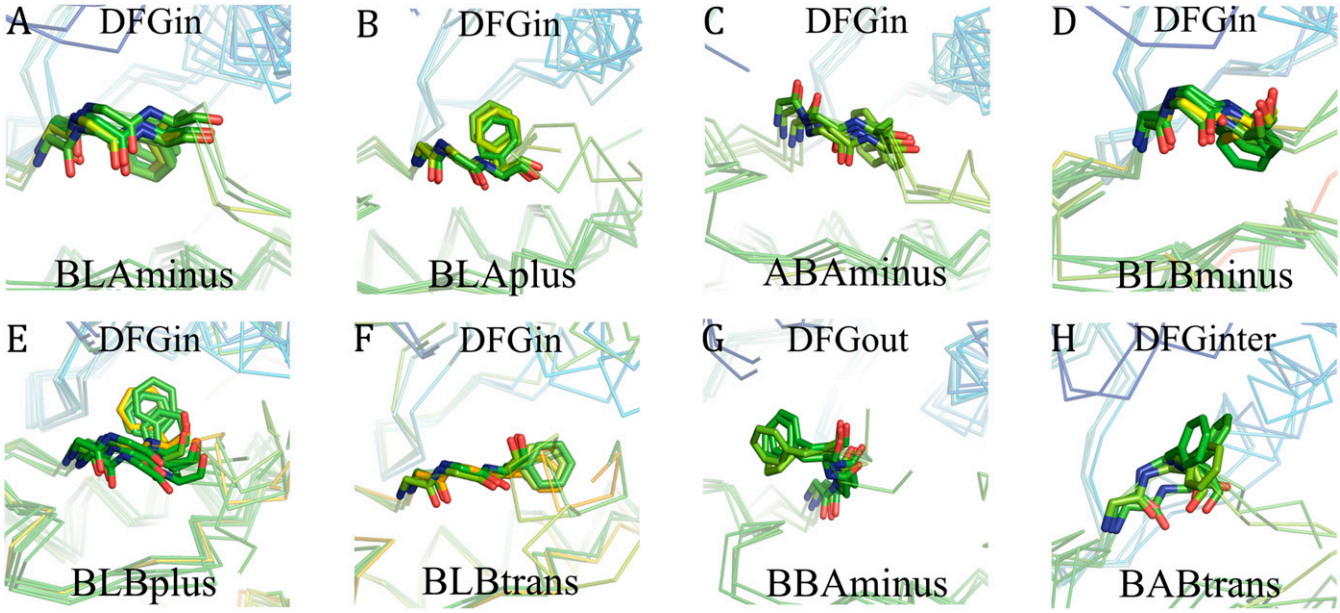

Fig. 5. Structure examples of each cluster. (*A*) BLAminus; (*B*) BLAplus; (*C*) ABAminus; (*D*) BLBminus; (*E*) BLBplus; (*F*) BLBtrans; (*G*) BBAminus (DFGout); (*H*) BABtrans (DFGinter).

residues or mutants thereof) are listed in Dataset S1. Mean dihedral angles and representative structures for each cluster are given in *SI Appendix*, Table S2.

Table 1 summarizes the features of each cluster. Kinase structures are most commonly observed in the BLAminus conformation (55.6% of 4,834 chains). The catalytically primed structures discussed earlier are a subset of this cluster. The next-most-frequent conformations are BLBplus and ABAminus observed in 9.5% and 9.1% of kinase chains, respectively. However, as discussed below, some of the structures in the ABAminus state are probably incorrectly modeled. A total of 4.1% of structures have a BLBtrans orientation but a large number of structures in this cluster are only from CDK2 kinase (160 of 199 chains). The remaining DFGin clusters are BLBminus (3.8) and BLAplus (3.1%). The two largest, inactive DFGin clusters are larger than the DFGout BBAminus cluster, which represents 5.1% of kinase chains. Although there are 113 chains (2.3% of all chains) observed in the DFGinter conformation, only 20 of them cluster into the BABtrans state (0.4%).

We examined the distribution of kinase domain sequences and kinase families in our eight clusters (domains in proteins with two kinase domains, such as the JAK kinases, are counted separately). The DFGin, DFGinter, and DFGout spatial groups are observed in the structures of 227, 27, and 68 kinase domain sequences, respectively. A total of 177 kinases with structures from all eight kinase families have been solved in the BLAminus conformation (Table 1). The other prominent DFGin clusters, ABAminus and BLBplus, were observed in 51 and 43 kinases, respectively. While all of the families have structures in the ABAminus cluster, BLBplus does not have structures from kinases in the CAMK or CK1 families. Forty-four kinases have structures solved in the DFGout BBAminus state; 25 of these are from the tyrosine kinase family. Among the eight subgroups, tyrosine kinase structures are the most diverse with a significant number of kinases having structures determined in all eight conformational states. Out of 244 human kinases with known structures, 187 have been observed in only one conformation.

Validation of Kinase Conformational Clusters with Electron Density. Our clusters depend on the precise directionality of the main-chain conformations of the X-D-F residues at the beginning of the activation loop of kinases. Segments of protein structures are sometimes incorrectly modeled within the electron density such that the peptide plane is flipped by 180°, resulting in large dihedral angle changes of neighboring residues (19, 20). Peptide flips change ψ of residue i and φ of residue $i + 1$ by about 180° while the other dihedrals remain approximately the same. In

Table 1. Number of human kinases in different conformations from each phylogenetic group

Spatial group	Clusters	All (n = 4,834)	Filtered (n = 1,621)	AGC (28)	CAMK (31)	CK1 (10)	CMGC (35)	STE (29)	TKL (18)	TYR (59)	Other (30)	Total (244)
DFGin	BLAminus	55.6%	58.6%	22	23	10	29	17	14	35	26	177
	BLAplus	3.1	2.4	2	1	0	1	3	1	7	2	17
	ABAminus	9.1	7.7	9	7	2	6	4	2	15	6	51
	BLBminus	3.8	5.0	0	3	0	4	4	3	8	4	26
	BLBplus	9.5	9.9	1	0	0	8	10	3	19	2	43
	BLBtrans	4.1	3.8	0	1	0	1	1	0	3	1	7
	Noise	4.4	2.4	4	5	0	9	9	4	22	11	64
DFGout	BBAminus	5.1	5.5	1	2	1	6	3	4	25	2	44
	Noise	2.9	2.7	4	1	0	4	7	2	20	4	42
DFGinter	BABtrans	0.4	0.2	0	0	0	0	1	0	6	1	8
	Noise	1.9	1.3	2	7	0	1	3	0	6	2	21

particular, we were concerned with some ABAminus structures that we observed to have poor electron density for the carbonyl oxygen atom of the X residue preceding the DFG motif (Fig. 6*A* and *SI Appendix*, Fig. S2). In Fig. 6*A*, the negative electron density contours of the Fo-Fc maps in red show that the O atom of the X residue has been modeled without electron density support, while the positive density in green indicates the positions where there is electron density but no atom has been modeled. Both of these features point to erroneous modeling of atoms in these structures.

We have quantified this error by using the EDIA (electron density score for individual atoms) program (21). EDIA computes the weighted mean of electron density over grid points in a sphere around the atom large enough to detect both unmodeled electron density and extra modeled density where an atom is placed incorrectly. A value of 0.8 or more reflects a good electron density fit; values lower than 0.8 indicate a problem in the model. The only atoms with average EDIA scores below 0.8, consistent with a large population of mismodeled structures, are the O atoms of the X residue of ABAminus structures (Fig. 6*B*) and the O atom of the Phe residues of BLAplus structures (*SI Appendix*, Fig. S3). This is consistent with a significant population of ABAminus structures that should have been modeled as the more common BLAminus conformation (180° change in ψ_X and φ_D) and BLAplus structures that should have been modeled as the much more common BLBplus conformation (180° change in ψ_F and φ_G).

A visual inspection of the electron density in mismodeled ABAminus structures suggested that the error in modeling is more commonly encountered in structures with a DFG-Asp χ_2 dihedral of about 0°. This value of χ_2 leads to a superposition of the Asp carboxylate in ABAminus and BLAminus structures. To examine this, for the ABAminus structures we plotted the EDIA score of the O atom of the X residue versus χ_2 of the Asp side chain (Fig. 6*C*). In correctly modeled ABAminus structures, the Asp χ_2 dihedral is around 90°, and the EDIA scores of ABAminus and BLAminus structures are approximately the same, while in structures with poor electron density it is around 0° with EDIA scores significantly lower (minimum average EDIA = 0.63) than BLAminus structures with similar dihedral angles (maximum average EDIA = 0.92).

With these results, we have produced a filtered dataset consisting of structures with resolution ≤2.25 Å, R-factor ≤0.25, and EDIA scores above 0.8 for all backbone atoms of the XDF motif (Dataset S2). All further analysis of the clusters was performed for this filtered dataset. Table 1 demonstrates that the frequency of ABAminus and BLAplus in the filtered set decreases as expected, as does the noise (from 9.2% to 6.4%).

Conformational Properties of Kinases with Specific Ligands. The distribution of ATP-bound and apo structures across our clusters may give an indication of what inactive conformations some kinases prefer in vivo, although we are limited by the structures determined for each kinase and the conditions employed during crystallization. Table 2 lists the percent of ATP-bound (or analog), inhibitor-bound, or apo structures in each of our clusters for the filtered dataset. Seventy-six percent of ATP-bound structures are in the BLAminus cluster. A triphosphate is never found in complex with a kinase in the DFGout and DFGinter states because the location of the DFG-Phe side chain would make the binding of ATP unfavorable. Among ATP+Mg–bound structures, only the BLAminus, BLBplus, and BLBtrans states are represented. As expected, there are no ABAminus structures with bound magnesium in the filtered dataset.

The apo-form kinases are predominantly in the BLAminus state (64%); the other DFGin structures comprise 25% of the apo structures. This shows that multiple clusters which are observed within the DFGin group are likely to be naturally occurring states even in the absence of any bound inhibitor. However, there are relatively few apo structures of kinases in the DFGinter cluster (BABtrans, 1.3% of apo structures) and none in the DFGout BBAminus cluster, and as noted above ATP-bound structures do not either. These distributions are in contrast to inhibitor-bound structures, which are observed across all of the conformations. A total of 239 chains (96%) in the BBAminus cluster are in complex with an inhibitor, 200 chains (80%) of which are in complex with a type II inhibitor (listed in Dataset S3).

Our clustering scheme can be applied to determine the conformational states bound by specific inhibitors (Fig. 7). For example, sunitinib is bound to several DFGin conformations: STK24, PAK6, and PHKG2 in BLAminus conformations; ITK in a BLAplus conformation; and CDK2 in a BLBminus conformation (Fig. 7*A*). It is also bound to tyrosine kinases KIT and VEGFR2 in DFGout-BBAminus conformations. Dasatinib binds to different conformations of the same kinase in different PDB entries: ABL1 in BLAminus and BABtrans states, and BTK in BLAplus and BABtrans conformations. It also binds to

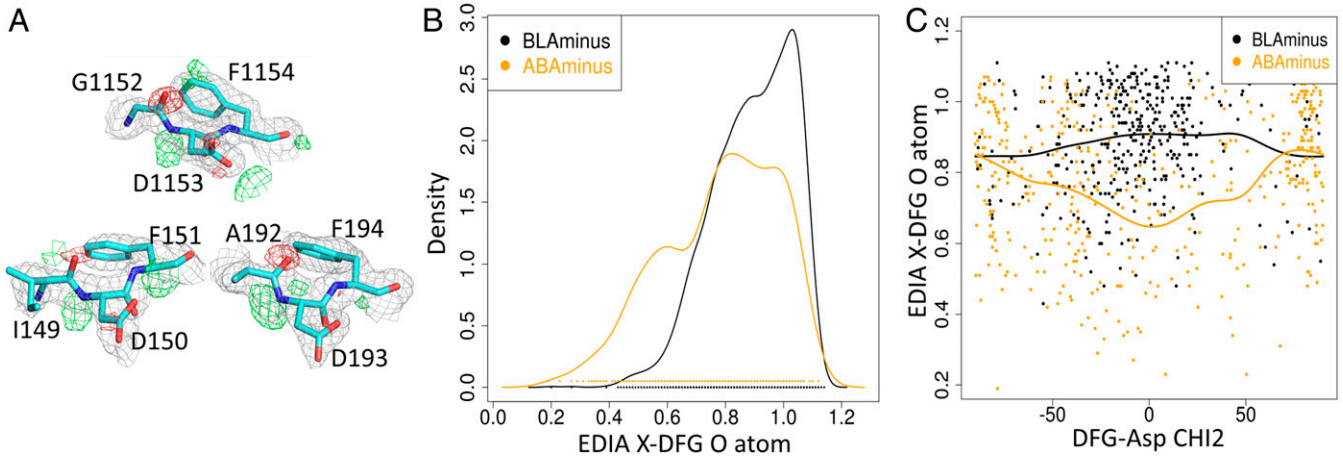

Fig. 6. Electron densities indicate incorrectly modeled ABAminus structures. (*A*) Electron density of ABAminus structures consistent with BLAminus structure. 2Fo-Fc (gray); Fo-Fc (green, positive density, indicating density not represented by an atom; red, negative density, indicating density where an atom has been placed without electron density support). PDB ID code 1K3A_A (IGF1R); 4BL1_A (MELK); 3IEC_D (MARK2). (*B*) Kernel density estimates of the EDIA scores of the backbone carbonyl atom (X_O) of the X-DFG residue for ABAminus (orange) and BLAminus (black) structures. (*C*) Scatterplot and von Mises kernel regressions of EDIA of X_O vs. χ_2 of DFG-Asp of ABAminus (orange) and BLAminus (black) structures. BLAminus structures mostly have χ_2 near 0°. ABAminus structures with χ_2 near 0° have poor electron density and may be mismodeled BLAminus structures.

Table 2. Distribution across clusters for different types of ligands

Cluster	Apo (n = 149)	ATP (118)	ATP+Mg (60)	Inhibitors (1,308)
BLAminus	64.4	76.2	68.3	58.7
BLAplus	4.0	—	—	2.5
ABAminus	6.7	1.6	—	8.8
BLBminus	4.0	—	—	5.8
BLBplus	5.3	5.9	11.6	11.3
BLBtrans	2.6	11.8	20.0	3.4
Noise	2.6	3.3	—	2.4
BBAminus	—	—	—	6.9
Noise	5.3	—	—	2.8
BABtrans	1.3	—	—	—
Noise	3.3	—	—	1.2
Total, %	100	100	100	100

ATP, triphosphates (ATP, ANP, ACP, and AGS); ATP+Mg, both triphosphate and Mg bound. Dashes indicate none.

BLAminus conformations of ABL2, PMYT1 SRC, EPHA4, STK10, and STK24 as well as BLBplus in PTK6, BLBminus in EPHA2, and DFGinter-BABtrans structures in BMX (Fig. 7*B*). Fig. 7*C* shows similar results for bosutinib. As a type II inhibitor, imatinib binds to DFGout structures in the BBAminus state (Fig. 7*D*). This kind of analysis makes clear that some inhibitors do not bind to all kinases in the same conformational state or even to one kinase in only one conformational state. It also argues that classifying inhibitors by the state of the kinase they bind to is not necessarily useful. It may be more productive to classify them by what volumes within the kinase active and C-helix sites they occupy (8).

What Is Wrong with Inactive Structures? The kinase active site requires a number of moving parts to be placed precisely for catalysis (Fig. 2*A*). While previous schemes have associated DFGin structures with the active form, we have explicitly divided DFGin structures into an active form—BLAminus—and five inactive forms. In addition, we have identified the most common DFGout form (BBAminus) and a group we call DFGinter (BABtrans). We examined several structural features across our clusters to determine what is commonly missing for activity in the inactive (non-BLAminus) clusters: (*i*) an orientation of the C-helix such that the C-helix–Glu/β3 strand–Lys salt bridge can form; the lysine side chain must be in a specific position to form a hydrogen bond with an oxygen atom of the α phosphate and the oxygen atom bridging the β and γ phosphates of ATP; (*ii*) the position and orientation of the Asp residue of the DFG motif; in catalytically primed structures, the Asp backbone NH positions a water molecule that forms hydrogen bonds with the adenine ring of ATP, while the carboxylate Oδ2 atom positions the Mg^{2+} ion which interacts with the β phosphate of ATP; (*iii*) an extended activation loop is required for binding of substrate to the kinase active site; in inactive structures, the activation loop is folded up, blocking access of the active site to substrates. The results of this analysis are presented in Table 3. We examine each of these three features in turn.

We define C-helix-in and C-helix-out structures (those that can or cannot form the Glu/Lys salt bridge) as those with a distance between the C-helix–Glu–Cβ and β3–Lys–Cβ atoms ≤10 Å or >10 Å, respectively. In structures with an intact Glu/Lys salt bridge, 98% of the Cβ–Cβ distances are less than 10 Å (*SI Appendix*, Fig. S4). This distance is therefore characteristic of the ability of the kinase to form the Glu-Lys salt bridge, regardless of whether the side-chain atom positions of these residues were resolved in the structure. In the BLAminus cluster in the filtered dataset, which includes the catalytically primed conformation of kinases, the C-helix is in an inward orientation in 95% of the structures (Lys/Glu in Table 3). Among the other DFGin clusters, ABAminus has the highest frequency of chains in a C-helix-in conformation (89%). By contrast, the large BLBplus cluster is strongly associated with a C-helix-out conformation (77%). In this cluster the g-plus rotamer of DFG-Phe points upward, pushing the C-helix outward. Surprisingly, in 75% of the DFGout structures in the BBAminus cluster, the C-helix remains in an inward disposition, suggesting that type II inhibitors do not push the helix outward despite occupying a pocket adjacent to the C-helix.

The position and orientation of the Asp carboxylate enable the chelation of magnesium ion in catalytically primed structures. The position of the Asp backbone and side chain can be assessed in part by the presence of a hydrogen bond between the carbonyl of the X-DFG residue and the histidine side chain of the HRD motif. When this hydrogen bond is present, the NH of Asp points upward so that it can form interactions with water molecules and the adenine ring. If the Asp side chain reaches the correct rotamer, it is able to form interactions with magnesium, which then interacts with ATP (Fig. 2*A*). We used the presence of the Asp-O/HRD-His-Nε2 hydrogen bond as a proxy for proper positioning of DFG-Asp (AspNH in Table 3). All of the clusters except ABAminus have this hydrogen bond in a majority of structures. Because the carbonyl of X-DFG points upward in ABAminus, the NH of DFG-Asp points downward and is unable to position water molecules that interact with the adenine moiety of ATP. As we observed in Fig. 6*C*, the aspartic acid side chain has a χ_2 dihedral angle of near 90° in most ABAminus structures with good electron density. In this orientation, DFG-Asp would not bind magnesium ion without a rotation of χ_2 which is unfavorable in this combination of backbone conformation and χ_1 rotamer (22).

An extended conformation of the activation loop is required for substrate binding to kinases, since folded loop structures block the substrate binding site. When the activation loop is extended, the backbone N atom of the sixth residue in the loop (DFGxx**X**) makes a hydrogen bond with the backbone oxygen atom of the residue preceding the HRD motif (X-HRD, Fig. 2*A*). Using this criterion, our analysis shows that 99% of chains in the BLAminus cluster have an extended activation loop in the filtered dataset (Loop in Table 3). Among the other DFGin

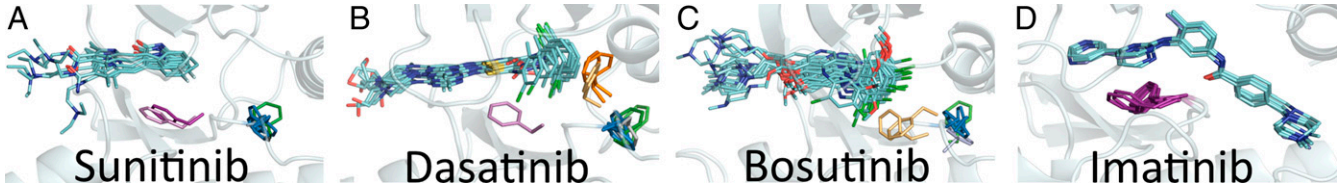

Fig. 7. Inhibitors bound to structures in multiple clusters: BLAminus (sky blue), BLAplus (dark green), BLBminus (light gray), BLBplus (light green), BBAminus (purple), BABtrans (orange), DFGout-noise (light purple), and DFGinter-noise (light orange). (*A*) Sunitinib bound to PAK6, PHKG2, and STK24 in BLAminus; ITK in BLAplus; CDK2 in BLBminus; KIT in BBAminus; KIT in DFGout-noise. (*B*) Dasatinib bound to ABL1, STK10, and STK24 in BLAminus; BTK in BLAplus; EPHA2 in BLBminus; PTK6 in BLBplus; BMX and BTK in BABtrans; ABL1 in DFGinter-noise; DDR1 in DFGout-noise. (*C*) Bosutinib bound to EPHA2, PMYT1, SRC in BLAminus; ERBB3 in BLBplus; STK10 in DFGin-noise; ABL1 and KCC2A in DFGinter-noise. (*D*) Imatinib bound to ABL1, ABL2, CSF1R, DDR1, KIT, and LCK in BBAminus.

Table 3. What is wrong with inactives?

Cluster	Lys/Glu	AspNH	Loop	All
BLAminus (951)	94.6	99.7	99.1	93.9
BLAplus (39)	71.7	97.4	—	—
ABAminus (126)	89.6	—	45.2	—
BLBminus (82)	48.7	100	1.2	1.2
BLBplus (161)	22.3	100	—	—
BLBtrans (62)	—	100	—	—
Noise (40)	45.0	50.0	2.5	2.5
BBAminus (90)	75.5	95.5	8.8	—
Noise (45)	75.5	71.1	6.6	—
BABtrans (4)	25.0	75.0	25.0	—
Noise (21)	52.3	61.9	—	—

Data from the filtered dataset. Lys/Glu, formation of salt bridge possible (Chelix in); AspNH, DFG-Asp-NH in right position to interact with water molecule that forms hydrogen bond with ATP, as measured by percent hbond of X-DFG carbonyl with HRD-His side chain; Loop, substrate binding possible because activation loop is extended; All, percent of structures where all three features are present.

clusters, 45% of ABAminus chains have their activation loop in a similar conformation. Beyond BLAminus and ABAminus, an extended activation loop is rare.

If we take these three criteria into account, only the BLAminus structures are capable of binding ATP, magnesium ion, and substrate simultaneously with 94% of structures passing all three (All in Table 3). The only structures from other clusters that pass the three criteria consist of one DFGin noise structure and one BLBminus structure, both from the same PDB entry, human IRAK4, PDB entry 2NRY, chains A and B, respectively. Those chains also exhibit poor electron density for the residues after DFG. Notably, the other two monomers in this crystal (chains C and D) are BLAminus structures.

To add some additional detail to the structural features of our clusters, we examined them for the side-chain conformation of the DFG-Asp side chain and the presence of beta turns (*SI Appendix*, Table S3), and the position of the Gly-rich loop (*SI Appendix*, Fig. S5). The DFG-Asp side chain has a predominant rotamer in each of our clusters. Ninety-six percent of BLAminus structures possess a type I beta turn beginning with DFG-Phe (sequence FGXX) with a hydrogen bond between the carbonyl of DFG-Phe and the NH of the DFG+2 residue. This beta turn determines the path of the activation loop into an extended conformation suitable for substrate binding. Fifty-five percent of ABAminus structures contain this beta turn but no other clusters do. The other clusters contain a beta turn beginning at DFG-Asp, including 60% of BLBplus structures that have a type II turn (sequence DFGX) and 49% of BLAplus structures with a type I turn at this position.

Comparison with Previous Kinase Classification Schemes. We compared our labels to three previously published classification schemes. Taylor and coworkers (23) have defined a regulatory spine as a stacking of four hydrophobic residues which dynamically assemble in the active state of the kinase. These consist of the HRD-His, the DFG Phe, C-helix Glu+4, and a residue in the loop just before strand β4 (Fig. 8). Taylor and coworkers (23) defined the regulatory spine only by comparison of structures; they did not define what constitutes the presence or absence of the spine in any one structure. We took a simple approach and define the regulatory spine as present if the minimum side-chain/side-chain contact distances among these four residues (1, 2, 2, 3, 3, 4) are all less than 4.5 Å. The regulatory spine is present in 98% of the BLAminus structures and 100% of the BLAplus structures, but it is also present in about 70% of BLAplus, ABAminus, BLBminus, and BLBplus clusters, indicating that its presence is not sufficient for defining active kinase structures (Fig. 8 and *SI Appendix*, Table S4). Rather, it is a feature of most DFGin structures, whether they are active or inactive enzymes. The regulatory spine is never intact in DFGout and DFGinter structures, because the DFG-Phe residue has moved out from the back pocket, making the first distance greater than 4.5 Å in all of these structures.

We compared our clusters with the Möbitz scheme (*SI Appendix*, Table S5) (11). The scheme uses DFGin and DFGout as the major labels and subdivides them into seven and five conformations, respectively. Four of the 12 are unintuitively named for some representative members of each cluster ("AuP BRAF," etc.). We found that although this classification is able to identify most of the BLAminus structures as active, it also incorrectly defines most of the ABAminus structures as active. It divides our BBAminus type-II-inhibitor binding structures into three DFGout labels. Further, it merges our DFGinter and DFGout-noise structures into a "DFG Flipped" category, even though the positions of the DFG-Phe residues in these two conformations are far apart.

We compared our classification with that of Ung et al. (12) (*SI Appendix*, Table S6), which divides kinases into CIDI, CODI, CIDO, CODO (C-helix in/out and DFGin/out), and a DFGinter-like group, ωCD. Ung et al.'s method successfully distinguishes DFGinter conformations, which they call ωCD. However, as with many previous schemes, it fails to distinguish active from inactive DFGin structures and lumps our six DFGin clusters into their CIDI and CODI clusters based solely on the position of the C-helix. However, as our clustering has shown, these conformations have different conformations of the activation loop (Fig. 5 *A*–*F* and Table 3).

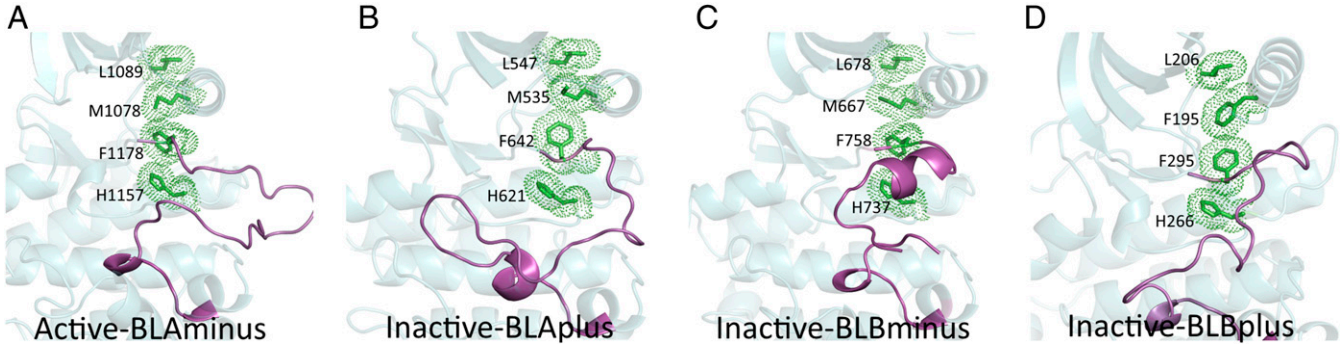

Fig. 8. The intact regulatory spine of active and inactive structures. (*A*) Active BLAminus (3BU5_A, INSR); (*B*) inactive BLAplus (4F64_A, FGFR1); (*C*) inactive BLBminus (5NK7_A, EPHA2); (*D*) inactive BLBplus (4UY9_B, M3K9). We consider the spine intact if there is at least one atom–atom contact (≤4.5 Å) between the side chains of each residue of the spine.

Discussion

We have developed a clustering and labeling scheme which first divides the kinase structures into three groups, based on the location of the DFG-Phe side chain, that are further clustered based on the orientation of the activation loop. To cluster the orientation of the activation loop, we have used the dihedral angles that determine the placement of the Phe side chain: the backbone dihedrals (φ, ψ) of the X-D-F residues and the first side-chain dihedral (χ_1) of the DFG-Phe residue. These are parameters used to define the conformation of any polypeptide chain. From this clustering, we have developed a simple nomenclature for kinase conformations that is intuitive and easily applied by structural biologists when they determine a new kinase structure. It is based on the region occupied by the XDF backbone dihedrals on the Ramachandran plot and the side-chain rotamer of DFG-Phe.

One of the most important results of our clustering is that it is able to identify several distinct states within the ensemble of active and inactive DFGin structures, which have usually been grouped together in previous clustering schemes (9, 10, 15). We have determined that the most frequently observed conformation, BLAminus, is also the active-state conformation of kinases. Catalytically primed structures, those containing bound ATP and Mg^{2+}/Mn^{2+} ion and a phosphorylated activation loop, are all members of the BLAminus cluster. We find that nearly all BLAminus structures have structural features consistent with an active kinase.

Among the inactive states in the DFGin group, BLBplus and ABAminus are the most frequent conformations with almost the same frequency at 9.5% and 9.1%, respectively. However, we observed that many structures with ABAminus conformations are likely to be incorrectly modeled. In these structures, the peptide group spanning the X and D residues of the X-DFG sequence is flipped such that the backbone carbonyl oxygen of the X-DFG residue is misplaced. This kind of error in structure determination is fairly common, in this case leading to BLAminus structures being incorrectly modeled as far less common ABAminus structures. Upon removing low-resolution and poorly determined structures, BLBplus becomes even more prevalent than ABAminus (10% and 7.7%, respectively) and is the most frequently occurring inactive conformation of kinases. In this conformation, the DFG-Phe ring is underneath the C-helix but pointing upward and the C-helix is pushed outward, creating extra volume, a region which is sometimes exploited for inhibitor design. BLBplus is sometimes referred to as the "SRC-like inactive" state (24, 25), although the latter has not been explicitly defined.

We have also examined why each type of inactive state is inactive. In the three BLB states (BLBplus, BLBminus, and BLBtrans), the C-helix is pushed outward in more than 50% of cases such that the Glu/Lys salt bridge in the N-terminal domain cannot form. In the ABAminus and DFGout and DFGinter states, the Asp side chain is not positioned to bind Mg so that it can interact with ATP. In all of the inactive states except ABAminus, the activation loop is not extended in a way that allows substrate binding.

We have compared our clustering and labeling scheme with three previously published methods. The regulatory spine defined by Taylor and coworkers (23) is a commonly used method to distinguish between active and inactive states, although it has not been explicitly defined. We find that the regulatory spine can only reliably distinguish DFGin structures from DFGout and DFGinter structures, failing to identify the different DFGin inactive states, most of which have an intact regulatory spine. Möbitz (11) developed a classification scheme that classified DFGin structures into seven states and DFGout structures into five states. There is a rough correspondence of our DFGin states to his, although our nomenclature is more intuitive and easier to apply by structural biologists. The scheme of Ung et al. (12) divides both the C-helix and DFG-Phe positions into in and out states, as have many previous schemes. As such, it does not capture the variability of our six DFGin clusters and fails to separate active from inactive kinase structures.

Inhibitors have been classified into type I and type II inhibitors, depending on whether they bind to the ATP binding site alone or to the ATP binding site and the volume next to the C helix that is exposed in DFGout structures. Our classification scheme allows us to determine that some inhibitors bind to different conformational states in different kinases, and sometimes to different states of the same kinase. In some cases, the inhibitor makes quite different contacts when bound to different states of the kinase, since the X-DFG residues may come in contact with the inhibitor. Some type I inhibitors even bind to DFGout conformations in our BBAminus cluster. We provide a list of structures of Food and Drug Administration–approved kinase inhibitors bound to kinases with cluster labels provided in *SI Appendix*, Table S7.

Our clustering and nomenclature can be applied to interpret the dynamical properties of various conformational states of kinases and the transitions between them. For instance, Tong et al. (26) studied SRC kinase with variants of dasatinib that stabilized three states of the kinase. In our nomenclature, these three states are BLAminus, BBAminus, and BLBplus. The BLBplus structure had significantly reduced dynamics of the HRD loop, the activation loop, and the loop between the F and G helices. Multiple studies have used molecular dynamics simulations to study the transition from active to inactive states in protein kinases (27–29). However, due to lack of a consistent nomenclature divergent DFGin states have been labeled as Src-like inactive. Levinson et al. (25) performed simulations of ABL1 from an "SRC-like inactive state" (PDB ID code 2G1T) that we label a BLBplus structure. However, Shan et al. (27) started simulations of EGFR from an "SRC-like inactive" structure (PDB ID code 2GS7), which is actually BLBtrans, an infrequently observed conformation for most kinases. Our scheme for assigning structures to different conformational states will improve the analysis of molecular dynamics simulations of kinases described in these studies.

Finally, significant effort has been expended to produce comparative models of kinases in different conformational states and to study the docking of inhibitors to these structures (30). A more reliable classification of the states of kinases will have a positive impact on choosing templates for producing models of kinases in various biologically and therapeutically relevant states.

Methods

The structures having kinase domains were identified from the file *pdbaa* (December 1, 2017) in the PISCES server (dunbrack.fccc.edu/pisces) with three rounds of PSI-BLAST (31) with the sequence of human Aurora A kinase (residues 125–391) as query. This profile was used again to search *pdbaa* with an E-value cutoff of 1.0×10^{-15} to eliminate structures of proteins that are not kinases or contain divergent folds. The structures with resolution worse than 4 Å and those with missing or mutated residues in the DFGmotif were removed. The conserved motifs were identified from pairwise sequence and structure alignments with Aurora A. Clustering was performed on all chains containing a kinase domain in these entries. We have defined a filtered dataset of high-resolution protein kinase structures. This set includes structures which satisfy the following criteria: (*i*) resolution better than or equal to 2.25 Å, R-factor ≤0.25, and no pseudokinases; (*ii*) EDIA score of X-DFG, DFG-Asp, and DFG-Phe backbone atoms ≥0.8; and (*iii*) overall EDIA score of DFG-Asp and DFG-Phe residues (including side chains) ≥0.8. The pseudokinases we removed from the dataset comprise CSKP, ERBB3, ILK, KSR2, MKL, STK40, STRAA, TRIB1, VRK3, WNK1, and WNK3. The filtered dataset is provided in Dataset S2.

To capture the location of the DFG-Phe residue, we calculated its distance from two conserved residues in the binding site: (*i*) αC-Glu(+4)-Cα to DFG-Phe-Cζ and (*ii*) β3-Lys-Cα to DFG-Phe-Cζ. We clustered structures into three groups using the sum of squares of these distances and average linkage hierarchical clustering algorithm using the hclust function in the statistical software R (32).

Dihedral angle clustering of the DFGin, DFGinter, and DFGout groups was performed with DBSCAN from the fpc package (33) in the R program. For

each pair of structures, *i* and *j*, the distance between them was calculated as the sum of an angular distance function over seven dihedral angles:

$$D(i,j)=\frac{1}{7}\left(d\left(\phi_i^X,\phi_j^X\right)+d\left(\Psi_i^X,\Psi_j^X\right)+d\left(\phi_i^D,\phi_j^D\right)+d\left(\Psi_i^D,\Psi_j^D\right)+d\left(\phi_i^F,\phi_j^F\right)+d\left(\Psi_i^F,\Psi_j^F\right)+d\left(X_{1i}^F,\chi_{1j}^F\right)\right).$$

The distance for each angle is equal to the square of the chord length between the ends of two vectors originating at the center of a unit circle (15, 16): $d(\theta_1,\theta_2)=2(1-\cos(\theta_2-\theta_1))$.

DBSCAN requires two parameters, ε and *MinPts*. Data points with at least *MinPts* points within a distance ε are considered "core points." Points within ε of a core point but not themselves core points are called border points. All other points are considered noise. If we make a graph by treating the core points as nodes and place edges between them if they are within ε of each other, then the clusters are identified as the connected subgraphs of the whole graph. Border points are then assigned to the cluster that contains the closest core point to the border point. The noise points are not assigned to clusters.

The choice of appropriate parameters is critical for determining the clusters. The value of *MinPts* is approximately equal to the smallest cluster that the procedure will return. If ε is too small, some clusters will be inappropriately subdivided into many small, dense clusters. If ε is too large, adjacent clusters may be merged. We have an advantage that protein φ, ψ dihedral angle pairs naturally cluster within the Ramachandran map. In addition, outliers are easily identified in forbidden regions of φ, ψ space. We used these features to scan through pairs of ε and *MinPts* values for each spatial group separately, minimizing the number of points identified as noise without visibly merging clusters representing distinct basins of the Ramachandran map. We used ε and *MinPts* values of 0.05 and 20 for DFGin, 0.06 and 20 for DFGout, and 0.3 and 15 for DFGinter. We validated our clusters with the silhouette metric (*SI Appendix*, Fig. S1). We also demonstrate that changing the ε and *MinPts* parameters for the DFGin group either splits our clusters unproductively (*SI Appendix*, Figs. S6 and S7) or merges clusters inappropriately (*SI Appendix*, Figs. S8 and S9).

To assign labels to as many data points as possible, the noise points whose distance from their nearest cluster centroid is less or equal to 0.3 units (equivalent to an average dihedral angle difference of 21°) were assigned to those respective clusters. The remaining noise points were still labeled with one of the three spatial group labels.

Hydrogen bond analysis was performed with HBPlus (34). The classification of beta turns into different turn types was done using a Python program by Maxim Shapovalov (https://github.com/sh-maxim/BetaTurn18) (35). The graphs were made using various plotting functions in the statistical package R. The molecular images were created using Pymol (https://pymol.org/2/). Electron densities were calculated with the program PHENIX (36). Electron densities were validated with the program EDIA (22).

We have created a simple tool to assign kinase structures in PDB format to our clusters (dunbrack.fccc.edu/kinasetool). A web database and server will be forthcoming. All data used for clustering are presented in Dataset S1. Scripts for processing the data and producing the figures are available from the authors.

ACKNOWLEDGMENTS. We thank Maxim Shapovalov for providing a program to identify beta turn types. V.M. thanks Fox Chase Cancer Center for an Elizabeth Knight Patterson postdoctoral fellowship. This work was funded by NIH Grants R01 GM084453 (to R.L.D.) and R35 GM122517 (to R.L.D.).

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298:1912–1934.
2. Taylor SS, Kornev AP (2011) Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem Sci* 36:65–77.
3. Huse M, Kuriyan J (2002) The conformational plasticity of protein kinases. *Cell* 109: 275–282.
4. van Linden OP, Kooistra AJ, Leurs R, de Esch IJ, de Graaf C (2014) KLIFS: A knowledge-based structural database to navigate kinase-ligand interaction space. *J Med Chem* 57:249–277.
5. Grossi V, et al. (2012) Sorafenib inhibits p38α activity in colorectal cancer cells and synergizes with the DFG-in inhibitor SB202190 to increase apoptotic response. *Cancer Biol Ther* 13:1471–1481.
6. Zhang J, Yang PL, Gray NS (2009) Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 9:28–39.
7. Johnson LN (2009) Protein kinase inhibitors: Contributions from structure to clinical compounds. *Q Rev Biophys* 42:1–40.
8. Zuccotto F, Ardini E, Casale E, Angiolini M (2010) Through the "gatekeeper door": Exploiting the active kinase conformation. *J Med Chem* 53:2681–2694.
9. Jacobs MD, Caron PR, Hare BJ (2008) Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: Structure of lck/imatinib complex. *Proteins* 70:1451–1460.
10. Brooijmans N, Chang YW, Mobilio D, Denny RA, Humblet C (2010) An enriched structural kinase database to enable kinome-wide structure-based analyses and drug discovery. *Protein Sci* 19:763–774.
11. Möbitz H (2015) The ABC of protein kinase conformations. *Biochim Biophys Acta* 1854:1555–1566.
12. Ung PM-U, Rahman R, Schlessinger A (2018) Redefining the protein kinase conformational space with machine learning. *Cell Chem Biol* 25:916–924.e2.
13. Chiu Y-Y, et al. (2013) KIDFamMap: A database of kinase-inhibitor-disease family maps for kinase inhibitor selectivity and binding mechanisms. *Nucleic Acids Res* 41: D430–D440.
14. Dodson CA, et al. (2010) Crystal structure of an Aurora-A mutant that mimics Aurora-B bound to MLN8054: Insights into selectivity and drug design. *Biochem J* 427:19–28.
15. Mardia K, Jupp P (2000) *Directional Statistics* (Wiley, London).
16. North B, Lehmann A, Dunbrack RL, Jr (2011) A new clustering of antibody CDR loop conformations. *J Mol Biol* 406:228–256.
17. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI, Portland, OR), pp 226–231.
18. Hollingsworth SA, Karplus PA (2010) A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts* 1:271–283.
19. Touw WG, Joosten RP, Vriend G (2015) Detection of trans-cis flips and peptide-plane flips in protein structures. *Acta Crystallogr D Biol Crystallogr* 71:1604–1614.
20. Keedy DA, Fraser JS, van den Bedem H (2015) Exposing hidden alternative backbone conformations in X-ray crystallography using qFit. *PLoS Comput Biol* 11:e1004507.
21. Meyder A, Nittinger E, Lange G, Klein R, Rarey M (2017) Estimating electron density support for individual atoms and molecular fragments in X-ray structures. *J Chem Inf Model* 57:2437–2447.
22. Shapovalov MV, Dunbrack RL, Jr (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19:844–858.
23. Kornev AP, Haste NM, Taylor SS, Eyck LF (2006) Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci USA* 103:17783–17788.
24. Xu W, Doshi A, Lei M, Eck MJ, Harrison SC (1999) Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell* 3:629–638.
25. Levinson NM, et al. (2006) A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol* 4:e144.
26. Tong M, et al. (2017) Survey of solution dynamics in Src kinase reveals allosteric cross talk between the ligand binding and regulatory sites. *Nat Commun* 8:2160.
27. Shan Y, Arkhipov A, Kim ET, Pan AC, Shaw DE (2013) Transitions to catalytically inactive conformations in EGFR kinase. *Proc Natl Acad Sci USA* 110:7270–7275.
28. Shukla D, Meng Y, Roux B, Pande VS (2014) Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat Commun* 5:3397.
29. Yang S, Banavali NK, Roux B (2009) Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Natl Acad Sci USA* 106:3776–3781.
30. Dixit A, Verkhivker GM (2012) Integrating ligand-based and protein-centric virtual screening of kinase inhibitors using ensembles of multiple protein kinase genes and conformations. *J Chem Inf Model* 52:2501–2515.
31. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
32. R Core Team (2015) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna).
33. Hennig C (2018) Package fpc: Flexible procedures for clustering, version 2.1-11.1 (University College London, London).
34. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 238:777–793.
35. Shapovalov M, Vucetic S, Dunbrack RL, Jr (2019) A new clustering and nomenclature for beta turns derived from high-resolution protein structures. bioRxiv:10.1101/390211. Preprint, posted August 13, 2018.
36. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66:213–221.