Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Identification of Binding Specificity-Determining Features in Protein Families

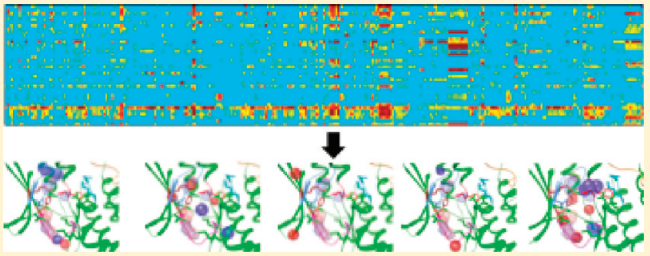
Peter C. Anderson,[†] Vincent De Sapio,[‡] Kevin B. Turner,[†] Sidney P. Elmer,[†] Diana C. Roe,[†] and Joseph S. Schoeniger*,[†]

[†]Sandia National Laboratories, Box 969, MS 9291, Livermore, California 94551, United States

[‡]HRL Laboratories, LLC., Information & Systems Sciences Laboratory, 3011 Malibu Canyon Road, Malibu, California 90265, United States

Supporting Information

ABSTRACT: We present a new approach for identifying features of ligand−protein binding interfaces that predict binding selectivity and demonstrate its effectiveness for predicting kinase inhibitor specificity. We analyzed a large set of human kinases and kinase inhibitors using clustering of experimentally determined inhibition constants (to define specificity classes of kinases and inhibitors) and virtual ligand docking (to extract structural and chemical features of the ligand−protein binding interfaces). We then used statistical methods to identify features characteristic of each class. Machine learning was employed to determine which combinations of characteristic features were predictive of class membership and to predict binding specificities and affinities of new compounds. Experiments showed predictions were 70% accurate. These results show that our method can automatically pinpoint on the three-dimensional binding interfaces pharmacophore-like features that act as "selectivity filters". The method is not restricted to kinases, requires no prior hypotheses about specific interactions, and can be applied to any protein families for which sets of structures and ligand binding data are available.



INTRODUCTION

The ability of small molecules and proteins to bind selectively is fundamental to the functioning of biological systems and is useful in numerous areas of chemical research, including drug design.[1] Improved understanding of the structural and chemical features that govern selectivity across large sets of proteins and ligands is needed to improve control of broad- and narrow-spectrum ligand binding, enable tuning of binding specificities, control off-target effects, and aid in the design of novel selective ligands and new protein receptors and enzymes. The selectivity of molecular recognition interactions can be defined properly only if many combinations of protein−ligand pairs are examined and differential binding correlated with structural features of the interactions. Although a wealth of structural and functional data is available for large, medically important families of proteins such as kinases[2−6] and viral proteases,[7,8] it is challenging for humans to synthesize information about thousands of binding affinities and millions of intermolecular binding contacts and form sufficiently rich hypotheses needed to predict molecular recognition specificity across these large target families.

In this work we present an automated method that identifies structural features that determine binding selectivity. Our computational pipeline employs a novel combination of binding-data clustering, virtual ligand docking, and statistical and bioinformatics analysis to yield three-dimensional models of specificity-determining features (SDFs) in binding interfaces. The identified structural features can be used to make predictions of ligand binding affinities and for ligand design.

Large-scale data sets of experimental ligand−protein binding interactions are becoming more numerous with improvements in experimental technologies and systems biology.[9] Previous studies using quantitative structure−activity relationship (QSAR) methods have demonstrated that important information about binding interactions can be revealed by clustering ligand−protein interactions based on molecular interaction fingerprints.[10−12] Proteins and ligands for which ligand−protein binding data exist can also be classified by clustering them according to their experimentally determined binding affinities, which requires no a priori knowledge of their bound complex structures.[13,14] Advances in genomics and structural biology have, however, produced a wealth of sequence and three-dimensional structure data for important protein classes,[15,16] posing a challenge of how to best integrate this wealth of data to formulate predictive hypotheses. Previous chemogenomic and proteochemometric studies that utilize large binding data sets have proven useful for modeling and predicting interactions and selectivity of receptors and enzymes with ligands and for the discovery of novel bioactive molecules.[13,14,17−30] However, these studies generally have not classified three-dimensional binding interface structures, involving both ligand and receptor structures simultaneously, but instead rely principally on sequence data and ligand or protein structure alone. Furthermore, they have not examined whole protein families to distinguish between subgroups for

Received: July 21, 2011
Published: January 30, 2012

specific sets of structural features in the ligand−protein binding interface itself that control ligand binding selectivity within the individual subgroups.

The interface that occurs between a protein and bound ligand is a complex surface containing numerous sites and diverse intermolecular interaction types.[31] There are frequently specific regions within the interface that contribute predominantly to binding energy and selectivity,[32,33] but they act in combination. Consequently, ligand binding is driven by the three-dimensional arrangement of binding site intermolecular interactions rather than by target amino acid sequence or ligand structure per se. Categorizing proteins according to their ligand binding preferences, and also ligands according to their protein binding preferences, would appear to require a *joint* classification of binding constant data *and* sets of binding interfaces. We hypothesized that such a joint classification would enable prediction of the binding selectivity of new proteins and ligands for which binding interfaces can be determined or modeled. We also hypothesized that, rather than classify the interfaces as complicated three-dimensional data objects, we could first classify ligands and proteins by binding data and extract from the simple chemical interaction features observed for all the complexes those interaction features that occur with high frequency in only some of the clusters. Machine learning could then discover combinations of these features that determine the similar binding behavior of the complexes within the cluster and differentiate them from other clusters. Also, if we preserve both chemical and structural information about the predictive features, the feature combinations can be useful for screening and to serve as templates for the design of new ligands and proteins.

Here, we examine the validity of our hypotheses by employing an extensive set of binding data for human protein kinases and a set of kinase inhibitors.[6] Kinases represent a good model for the classification of binding interfaces not only because they are structurally well characterized and extensively studied but also because they are an attractive target for therapeutic intervention in cancer,[34−36] inflammation,[37−42] diabetes,[43,44] arthritis,[45] neurodegenerative disorders,[46] and infection by HIV-1 and other pathogens that acquire drug resistance.[47] By combining virtual docking, proteochemometrics, bioinformatics, and statistical and experimental approaches, we implement an unsupervised pipeline to classify binding interactions across a large set of proteins and ligands and to identify SDFs at ligand−protein binding interfaces. SDFs are constellations of points within the ligand−protein binding interface where intermolecular hydrogen bonding interactions, polar−polar contacts, and hydrophobic interactions are formed. They consist not merely of specific protein residues but of specific spatial locations in the binding interface that function as hot spots for intermolecular interactions. As such, the SDFs are not necessarily sequence-specific and may apply to groups of proteins whose active sites have some degree of sequence variation.

We first establish that kinases can be clustered into distinct groups based on similarities of their binding affinity (K_d) profiles for a given set of inhibitors and, conversely, that we can cluster inhibitors based on similarities of their binding affinities for a given set of kinases. This finding is similar to earlier clustering studies[13,14,20] which use metrics other than K_d (IC_{50} values and percent inhibition) to compare protein and ligand binding profiles. Second, we show that within individual kinase clusters we can identify SDFs distributed over three-dimensional binding interfaces that are specific to the kinases within the clusters. These cluster-specific SDFs underlie the binding profile similarities of members within a given kinase cluster. As a validating example, the method is able to detect that the gatekeeper residue is a hydrophobic SDF for MAP kinases, in agreement with the gatekeeper hypothesis.[48,49] Furthermore, it is demonstrated that the identified SDFs can be used in conjunction with machine-learning methods to induce models for accurately predicting kinase inhibition activity of compounds from outside the standard data set, as confirmed experimentally.

The reported findings have allowed us to develop intuitive, easily visualized, and predictive models for understanding the basis of binding selectivity of proteins and ligands. Kinases were selected as a proof-of-principle model because they are well characterized experimentally and structurally, and abundant data are available pertaining to the effects of small structural changes on ligand binding properties. The unsupervised pipeline presented here is equally applicable to other large-scale sets of proteins and ligands for which there exist tables of binding data and protein structures that can be aligned. Therefore, it is a tool that can exploit the vast amounts of structural and binding data that emerging fields such as chemogenomics have recently made available[50] in order to provide for better ligand binding selectivity.

RESULTS

Clustering Kinases and Kinase Inhibitors According to Binding Affinities. An interest in determining whether the table of experimental kinase-inhibitor binding data published by Karaman and co-workers[6] can be ordered to show patterns of binding selectivity motivated us to cluster the table data based on binding affinities. The dissociation constants within the table of binding data, which contains 317 kinase structures and 38 inhibitors, span a wide range from ∼10 pM to 10 μM (Figure 1). Chemical structures of the 38 inhibitors are shown in Figure 2. As demonstrated in Figure 3A, prior to clustering, a raw heat map of the 317 × 317 matrix of pairwise Euclidian

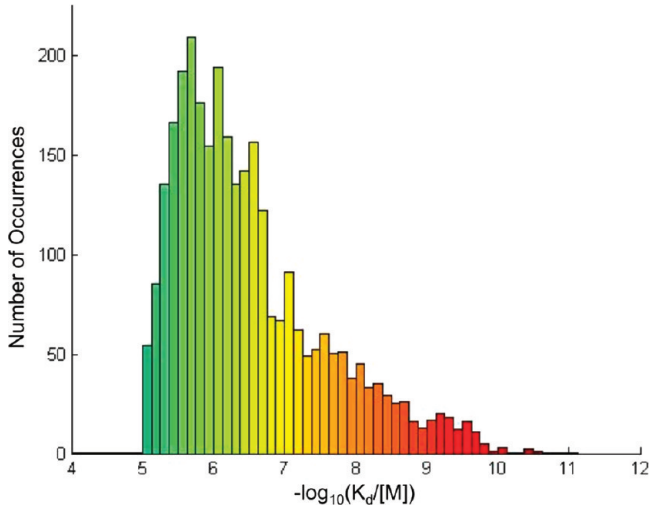


Figure 1. Histogram of experimental binding affinities (pK_d) of 38 inhibitors to 317 human kinases. Inhibitors with observed $pK_d < 5$ were considered nonbinders, in accordance with the work of Karaman et al.[6] Binding affinities are color-coded from weakest (aqua-blue) to strongest (dark red).

dx.doi.org/10.1021/jm200979x | J. Med. Chem. 2012, 55, 1926−1939

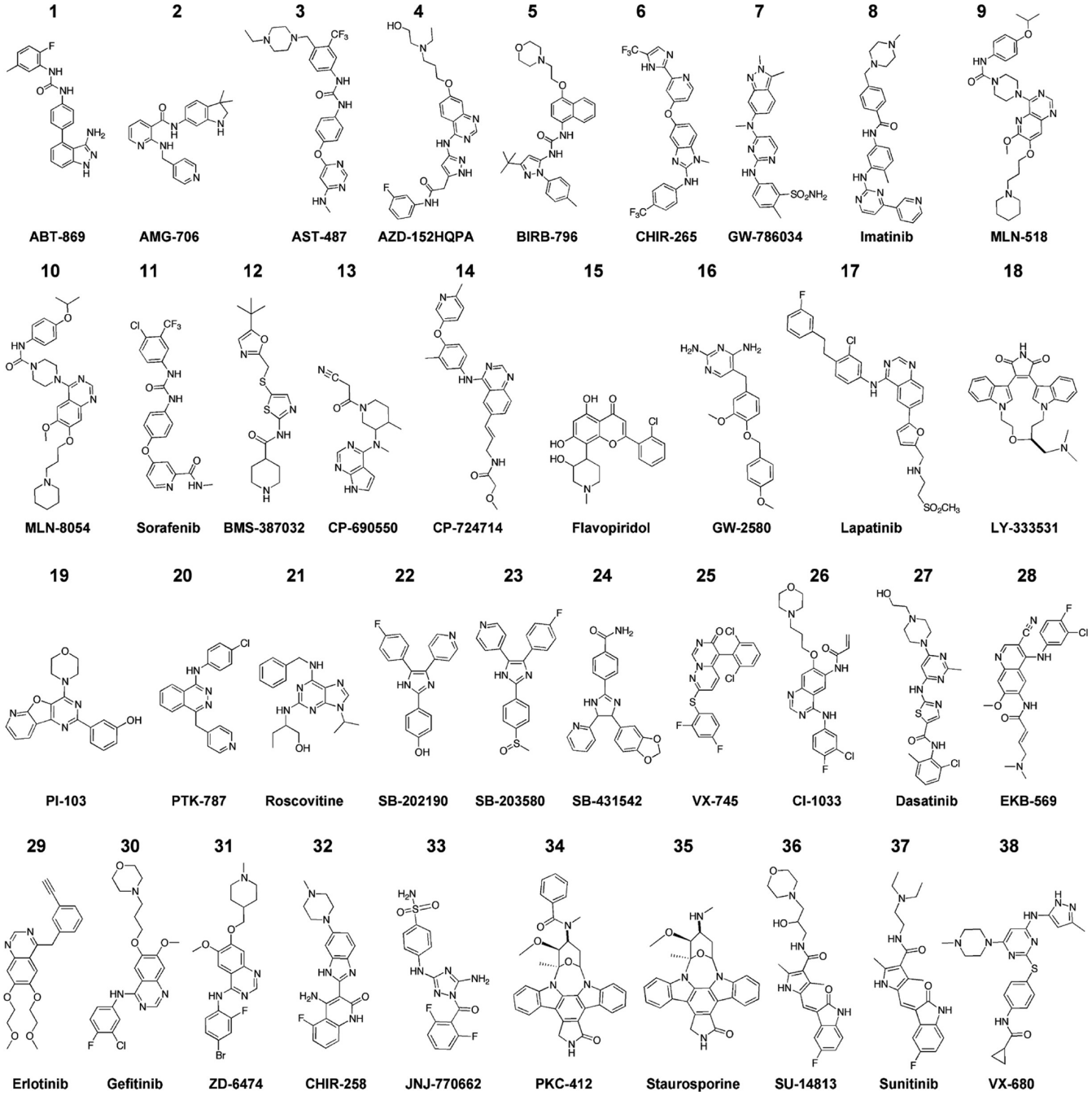


Figure 2. Structures of 38 kinase inhibitors used in present study. Numbers correspond to the inhibitor identification numbers in Figure 4B,C.

distances for the kinase binding affinity vectors shows no notable grouping. Eight kinase clusters were assigned in order to mimic the number of clusters reported in the human kinome phylogenetic tree,[51] leading to clear grouping following *k*-centers clustering (Figure 3B). Kinase pairs clustered within the squares centered on the diagonal in Figure 3B have highly similar binding profiles to the set of 38 ligands. Kinase clusters I–VIII contain 23, 11, 29, 83, 42, 64, 21, and 44 kinase structures, respectively. The members of each kinase cluster are listed in Table S1, Supporting Information. Similarly, clustering of the 38 × 38 matrix of pairwise Euclidian distances for the inhibitor binding affinity vectors results in clear grouping (Figure 3C). Optimum ligand clustering results based on consistency of cluster assignment during leave-one-out analysis were obtained when four ligand clusters were assigned. Ligand pairs clustered within the squares centered on the diagonal in Figure 3C likewise have similar binding profiles to the set of 317 kinases. Ligand clusters 1–4 contain 11, 14, 6, and 7 ligands, respectively, which are listed in Table S3, Supporting Information.

In order to demonstrate the value of clustering kinases and ligands together based on binding affinities, we plotted a heat map of the original, nonclustered 317 × 38 ligand–protein binding matrix (Figure 4A). This heat map shows no discernible binding patterns. When the binding matrix is ordered according to the four identified ligand clusters shown in Figure 3C, there is only a slight change in the heat map organization, producing no clear clustering (Figure 4B).

dx.doi.org/10.1021/jm200979x | *J. Med. Chem.* 2012, 55, 1926–1939

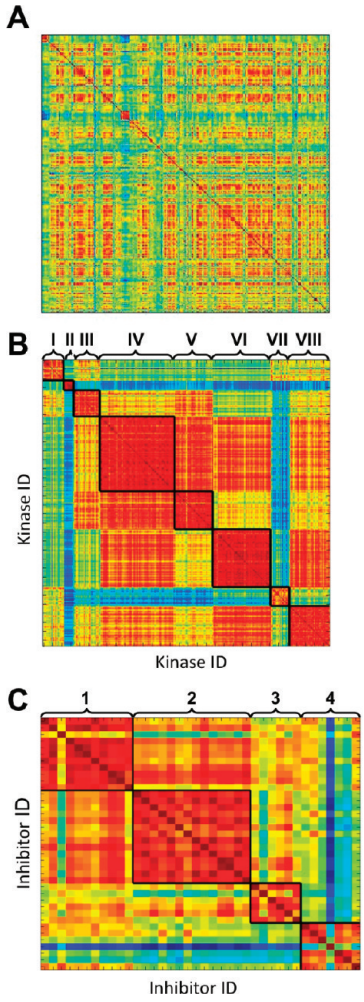


Figure 3. Heat maps of binding affinity similarities between kinases and between inhibitors. Pairwise Euclidian distances based on 38-element vectors of K_d values (for kinase–kinase distances) and the 317-element vectors of K_d values (for inhibitor–inhibitor distances) are plotted. Pairwise distances are represented by colors, with red and blue indicating smallest and greatest distances, respectively. (A) Unordered heat map of 317 × 317 kinase–kinase pairwise distances. (B) Ordered heat map showing *k*-centers clustering of the kinase–kinase pairwise distances. Eight clusters were assigned to mimic the number of clusters in the human kinase phylogenetic tree. Kinase cluster numbers are indicated by Roman numerals. (C) Ordered heat map showing *k*-centers clustering of the inhibitor–inhibitor pairwise distances. Optimum clustering occurred with the assignment of four clusters. Ligand cluster numbers are indicated with Arabic numerals.

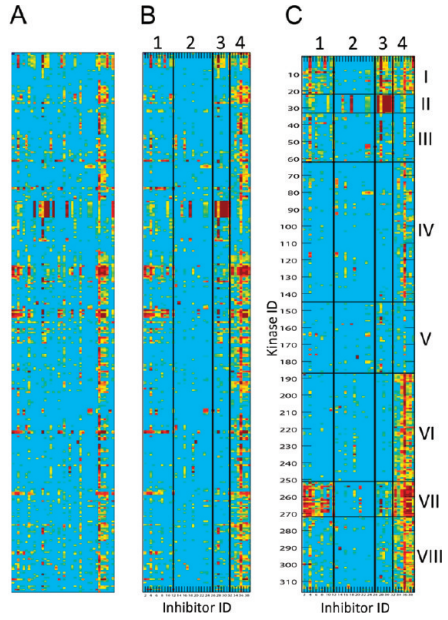


Figure 4. Heat maps showing binding matrix of experimentally measured binding affinities of inhibitors to kinases. Colors correspond to the binding affinities that are color-coded in the histogram of Figure 1. The background cyan color represents nonbinders (defined as K_d > 10 μM). (A) Unordered binding matrix. (B) Binding matrix ordered by four ligand clusters. Ligand clusters are numbered horizontally. (C) Binding matrix ordered by four ligand clusters (numbered horizontally) and eight kinase clusters (numbered vertically).

However, when the binding matrix is ordered according to both the identified ligand clusters and the identified protein clusters shown in Figure 3B, distinct groupings and binding patterns become apparent (Figure 4C). Kinases (or ligands) of the same cluster are located contiguously in the grid. The results indicate that kinases within a cluster tend to share strong similarity in the binding metric relative to each other and weaker similarities in the binding metric relative to kinases in other clusters.

The kinase and inhibitor clusters shown in Figure 4C have certain defining characteristics. Inhibitors in ligand cluster 1 bind with generally high affinity to the proteins in kinase cluster VII and with moderate affinity to many proteins in kinase cluster I, but, with the exception of AST-487, do not bind widely to any other kinase clusters. Ligand cluster 2 is sparse, and its members do not bind broadly to the members of any kinase cluster. Ligand cluster 3 is likewise sparse, with the exception that its members bind tightly to the proteins in kinase cluster II. Ligand cluster 4, in contrast, shows generally broad binding. Its members bind particularly broadly and tightly to the kinases in kinase clusters VI–VIII. Staurosporine, a member of ligand cluster 4, binds to the greatest number of kinases, including the majority of members of each of the eight kinase clusters, and has K_d < 10 μM for 288 of the 317 kinases in the data set. Ligand cluster 4 contains several other broadly binding inhibitors, including sorafenib, VX-680, SU-14813, and sunitinib, which bind with K_d < 10 μM to 85, 153, 182, and 198 separate kinase structures, respectively.

It is noteworthy that most, though not all, of the promiscuously binding compounds are clustered in ligand cluster 4. The clustering procedure does not attempt directly to group compounds into clusters of promiscuous and selective binders per se. Rather, the clustering procedure groups them according to the total Euclidean distances between their 317-member binding affinity vectors so as to place together ligands whose vectors have smaller Euclidean distances between them. Though AST-487 is a broad binder like staurosporine and sunitinib, it is assigned to cluster 1 because the wide range of kinases to which it binds has less overlap (average of 171 kinases in common) with those of the cluster 4 ligands than do the cluster 4 ligands with one another (average of 221 kinases in common). This smaller overlap yields a decreased Euclidean distance between AST-487 and the other ligands of cluster 1, leading to its assignment to ligand cluster 1.

Cluster Validation. For a description of the architecture of kinase active sites, the reader is referred to a detailed review by Liao.[52] Out of the 317 kinases in the table of binding data, 112 had experimentally available structures as of February 2010. An

all-versus-all comparison of the binding site structures of this subset of 112 kinases (listed in Table S2, Supporting Information) was performed by using the PocketMatch algorithm.[53] For each of the eight kinase clusters, an average score for all pairs of kinases belonging to the cluster (intracluster PM*Score*) was calculated. An average score for all kinases within a cluster matched with all kinases outside the cluster (extracluster PM*Score*) was likewise calculated. Higher scores indicate greater sequence and chemical similarity. The pairs of average PM*Scores* (intracluster, extracluster) for kinase clusters I–VIII are (42,34), (55,33), (37,32), (27,30), (35,33), (37,32), (37,31), and (37,35), respectively. The differences between each cluster's intra- and extracluster PM*Scores* are statistically significant ($P < 0.05$). In the case of seven of the eight kinase clusters, the average intracluster pairwise PM*Score* exceeds the average extracluster pairwise PM*Score*. This greater intracluster score indicates that kinases within a specific cluster tend to have binding sites that are structurally and chemically more similar to one another than to the binding sites of the kinases outside the cluster. Only cluster IV, which contains the largest number of members, has a larger intracluster PM*Score* than intercluster PM*Score*, demonstrating that there exists as much variability between its component binding sites as there exists across the entire kinase set. Figure S1 of the Supporting Information depicts the overlap of binding site structures for selected clusters.

These results indicate that structural differences between ligand binding sites produce distinct and recognizable patterns in ligand binding profiles. The results also establish that the kinase clustering, though it is based on experimental binding affinities of kinase inhibitors and not on structural information, nonetheless captures certain structural and chemical differences in ligand binding sites.

Kinase clusters were also compared to one another in terms of both full-length kinase domain and binding-site amino acid sequence-alignment scores (pairwise percentage sequence identity) using the ClustalW alignment program.[54−56] Average pairwise intracluster and extracluster alignment scores were computed for each of the eight kinase clusters, where extracluster scores are those between members of a specific cluster and all kinases outside the cluster. For full-length kinase domain sequences, the pairs of average sequence alignment scores (intracluster, extracluster) for clusters I–VIII are (46,21), (70,20), (33,19), (19,18), (17,17), (26,18), (39,20), and (16,18), respectively. The average binding-site sequence alignment scores for clusters I–VIII are (46,31), (47,25), (38,27), (30,29), (29,29), (31,27), (37,31), and (29,29), respectively. Clusters I, II, III, and VII have average intracluster alignment scores for both full-length and binding site sequences that are considerably greater than the corresponding extracluster scores. These four clusters are thus characterized by kinase members that have a higher degree of sequence identity with other kinases in the same cluster than with the kinase set as a whole.

Extracting Structural Features That Drive Clustering. *Virtual Docking To Predict Binding Conformations of Inhibitor–Kinase Complexes.* Because experimental structures of complexes are lacking, we used virtual docking to generate the binding interfaces for the complexes between the 38 kinase inhibitors and each of the 112 kinases with available crystal structures. All docking calculations were performed using AutoDock 4.2.[57] In order to estimate the accuracy of our docking protocol, we tested its ability to reproduce the experimentally determined conformations of a set of 75 ligands (5 of which are among our standard set of 38 inhibitors) bound to 104 kinase structures in a set of self-docking experiments. For 19%, 45%, and 75% of these 104 ligand–kinase complexes, the rmsd of the ligand heavy atoms from at least one of the two top-scoring poses relative to the experimental ligand conformation is ≤0.5 Å, ≤ 1.0 Å and ≤2.0 Å, respectively. This accuracy rate in self-docking compares well to that reported for other kinase docking studies.[58,59] In order to further evaluate the docking protocol, we performed cross-docking runs for the 38 inhibitors on a subset of 44 kinases. In 26% of all cases of an inhibitor docked to two separate kinase structures, the heavy-atom rmsd between the two top-scoring poses in each receptor structure is ≤2.0 Å. This percentage compares well with other cross-docking studies employing rigid receptors.[60,61]

Identification of Interaction Hot Spots Common to All Clusters. When we ran our spatial binning procedure (see Computational and Experimental Methods, Supporting Information), we found that there exist specific spatially localized intermolecular interactions that are shared predominantly by protein members belonging to single kinase clusters. In this paper, spatial bins that are highly populated with atoms participating in a specific intermolecular interaction type are referred to as *interfacial features*. Interfacial features were identified for intermolecular hydrogen bonds, hydrophobic interactions, and polar–polar contacts. We identified two classes of interfacial features: *global* interfacial features, which are shared by all kinase (or ligand) clusters, and *cluster-specific* interfacial features, which are ≥3 times more likely to be occupied among the complexes formed by the protein (ligand) members of a given kinase (ligand) cluster than among the complexes formed by any other cluster's members. We also designated whether interfacial features occur in *protein space* or *ligand space*. Interfacial features in protein space are bins that are highly populated by protein atoms participating in intermolecular interactions. Interfacial features in ligand space are bins that are highly populated by ligand atoms participating in intermolecular interactions (see Computational and Experimental Methods, Supporting Information, for further details).

Figure 5 shows the global interfacial features that are shared among kinases belonging to all eight kinase clusters. These features are the most frequently observed locations of protein atoms and ligand atoms participating in hydrogen bonding interactions, polar–polar contacts, and hydrophobic interactions across the set of 112 kinases and 38 inhibitors. Table 1 provides a summary of the frequencies of the type, class, and space (protein space and ligand space) of intermolecular interactions found at the sites of global interfacial features.

For the 8512 total docking poses (two top-scoring conformations for each of the 4256 inhibitor–kinase pairs), 13405 total hydrogen bonds are observed between inhibitors and kinase binding site residues. Two regions (Region 1 and Region 2 in Figure 5A), each comprising several spatially contiguous, globally shared interfacial features in ligand space, are observed for hydrogen bonding interactions. Together, regions 1 and 2 contain 4258 instances of ligand atoms participating in hydrogen bonding with the protein, corresponding to 32% of all observed hydrogen bonds. Of these 4258 instances of hydrogen bonding atoms, 2939 are hydrogen bond acceptors and 1319 are hydrogen bond donors. Ligand space region 1 is adjacent to the first three residues of the kinase hinge region (colored magenta in Figure 5) connecting the N-terminal and C-terminal kinase lobes. Region 1 accounts for 2820 (96%) of the 2939 ligand hydrogen-bond-accepting

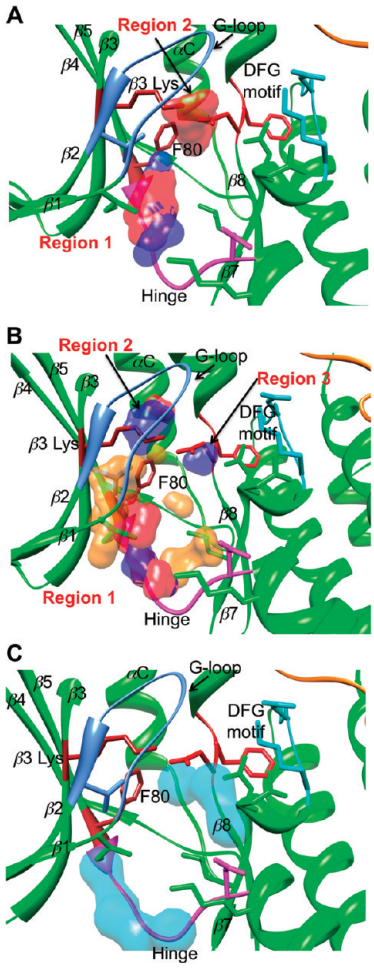


Figure 5. Ligand–protein interface features that are globally shared across all eight kinase clusters. For consistency with other figures, the structure of CDK2 (PDB code: 1B38) is depicted in all panels. Surfaces enclose contiguous features. (A) Most frequent locations for ligand hydrogen bonding atoms (ligand space). (B) Most frequent locations for protein hydrogen bonding atoms (protein space). Red and blue surfaces correspond to hydrogen bond acceptors and donors, respectively. Orange surfaces denote most frequent locations for hydrophobic interactions. (C) Most frequent locations for polar–polar contacts for protein atoms (protein space).

Table 1. Numbers of Instances of Intermolecular Interaction Types Occurring in Regions of Globally Shared Features and Cluster-Specific Features in Protein Space and Ligand Space[a]

	globally shared		cluster-specific		total	
	protein space	ligand space	protein space	ligand space	protein space	ligand space
HB donor	2659	1319	2227	1651	8523	4882
HB acceptor	1112	2939	1684	2247	4882	8523
polar–polar	31968	28460	6792	1578	80497	80497
hydrophobic	5876	NA[b]	3915	NA	25279	NA

[a]Numbers show total counts among the two top-scoring poses for all of the 4256 simulated ligand–protein complexes. [b]Hydrophobic interactions were characterized only for protein residues.

atoms that occur in regions 1 and 2 and for all the hydrogen-bond-donating atoms. Ligand space region 2 is situated next to the β3 Lys side chain (Lys33 in CDK2) and the Asp side chain of the DFG motif. This region comprises 119 (4%) of the ligand hydrogen-bond-accepting atoms.

Similarly, three regions consisting of spatially contiguous, globally shared hydrogen bonding interfacial features in protein space are observed (Figure 5B). Hydrogen bonding protein atoms within these regions interact primarily with those ligand atoms located within the ligand-space interfacial features shown in Figure 5A. Together, protein space regions 1–3 contain 3771 instances of protein atoms participating in hydrogen bonding with docked ligands, accounting for 28% of all observed hydrogen bonds. Of these 3771 hydrogen bonding atoms, 1112 are hydrogen bond acceptors and 2659 are hydrogen bond donors. Protein space region 1 contains the protein atoms that form hydrogen bonds with atoms in ligand space region 1 and comprises the first three residues of the kinase hinge region. The hydrogen bond donors and acceptors include primarily the main-chain amide nitrogen and oxygen atoms, respectively. Region 1 has 2855 instances of hydrogen bonding protein atoms, of which 1016 are hydrogen bond acceptors and 1839 are hydrogen bond donors. Protein space regions 2 and 3 comprise residues that donate hydrogen bonds to the acceptors in ligand space region 2. Protein space region 2 contains the β3 Lys side chain and has 623 hydrogen bonding atoms, while protein space region 3 contains the main-chain amide nitrogen of the DFG motif's Asp and Phe residues and has 293 hydrogen bonding atoms. The preponderance of hydrogen bonding interactions that we observe in the first three residues of the kinase hinge region agrees with other observations that these residues are the principal hydrogen-bond providers for ATP and inhibitor binding.[52]

Moreover, we calculated the most frequent locations of protein residues across all kinase clusters that take part in hydrophobic interactions with the docked inhibitors (represented as orange surfaces in Figure 5B). These hydrophobic interactions involve principally the following residues: a Val, Leu, or Ile residue immediately before the G-loop; a Val residue near the N-terminus of β2; the gatekeeper residue side chain (Phe80 in CDK2); a Leu or Val residue near the center of the hinge region; a Leu residue near the C-terminus of β7; a Val, Leu, or Ile residue in the αC-β4 loop or at the end of β8.

The most frequently occurring location of polar–polar ligand–protein contacts across all kinase clusters is positioned along the hinge region and near the Asp residue of the DFG motif (Figure 5C). This partial localization of polar–polar contacts along the hinge reflects the high prevalence of hydrogen bonding interactions in this region.

Several kinase inhibitors, including quercetin, PD98059, U0126, and BMS-509744, were selected from outside the standard set of 38 inhibitors and were docked to kinases from each of the eight kinase clusters. The top-scoring binding conformations of these docked complexes have the majority of their hydrogen-bond-donating and hydrogen-bond-accepting atoms positioned within the globally shared hydrogen bonding regions (data not shown). This observation suggests that the identified global hydrogen bonding interfacial features are not necessarily specific to the standard set of 38 inhibitors but can be generalized to inhibitors outside the standard set.

Identification of Cluster-Specific Interfacial Features. Each kinase cluster is characterized by a set of cluster-specific hydrogen bonding, polar–polar, and hydrophobic interfacial features. Cluster-specific features were examined in order to ensure that they are spatially distinct from global features of the same type.

dx.doi.org/10.1021/jm200979x | J. Med. Chem. 2012, 55, 1926–1939

Table 2. Total Numbers of Instances of Intermolecular Interaction Types Occurring at Locations of Cluster-Specific Features for Kinase Clusters I–VIII

protein space	I	II	III	IV	V	VI	VII	VIII
HB donor	159	29	0	64	61	37	1696	181
HB acceptor	287	43	22	60	127	269	536	340
polar–polar	672	674	1250	276	874	1054	1019	973
hydrophobic	511	185	421	425	519	901	318	635
ligand space	I	II	III	IV	V	VI	VII	VIII
HB donor	217	46	10	83	127	274	533	361
HB acceptor	180	28	0	69	59	39	1690	182
polar–polar	340	228	93	316	67	97	369	68

Table 2 lists the frequencies of intermolecular interactions occurring at locations of cluster-specific features across all analyzed ligand–protein complexes. The number of occurrences of atoms participating in intermolecular interactions within bins corresponding to cluster-specific features varies widely across the eight kinase clusters. For example, within the set composed of the two top-scoring poses for all 4256 modeled ligand–protein complexes, there are 1696 instances of hydrogen-bond-donating protein atoms located in the bins that correspond to the protein-space hydrogen-bond-donor features specific to kinase cluster VII. Correspondingly, there are 1690 instances of hydrogen-bond-accepting ligand atoms positioned in bins that correspond to the ligand-space hydrogen-bond-acceptor features specific to kinase cluster VII. In contrast, there are only 22 instances where hydrogen-bond-accepting protein atoms are situated in bins corresponding to protein-space hydrogen-bond-acceptor features that are specific to kinase cluster III.

Table S5 in Supporting Information presents data on the degree of uniqueness of the cluster-specific features for each intermolecular interaction type in protein space for the eight kinase clusters. The columns for kinase cluster numbers I–VIII indicate the relative proportions of ligand–protein complexes within that cluster where ≥1 protein atoms participate in the given intermolecular interaction type at the location of cluster-specific features of the cluster numbers listed in the rows. Proportions are scaled relative to the proportion for the kinase cluster corresponding to the row number. The table shows, for instance, that the bins of the hydrogen-bond-donor features that are specific to kinase cluster I are not populated by hydrogen-bond-donating atoms in any of the complexes formed by proteins belonging to clusters II–VIII. This uniqueness of occurrences in a single cluster is also the case for hydrogen-bond-donor features for clusters III, IV, and VI and for hydrogen-bond-acceptor features for clusters IV, V, and VI. In most cases, however, the features of a cluster are not uniquely populated by atoms from the cluster's own kinase members. Nevertheless, the relative probabilities that they are populated by interaction-forming atoms of kinases from other clusters are low. For instance, the proportion of complexes in kinase cluster I where a protein atom forms a hydrogen-bond-donating interaction within the bins of the hydrogen-bond-donor features specific to cluster II is ∼5 times smaller than the proportion for cluster II complexes themselves (0.22 vs 1 in the second row of Table S5). Similarly, the proportion of cluster III complexes where a protein atom forms a hydrogen-bond-donating interaction within the bins of the hydrogen-bond-donor interfacial features specific to cluster VIII is 50 times smaller than that of cluster VIII complexes themselves (0.02 vs 1 in the eighth row of Table S5).

Table S6 in Supporting Information shows data about the uniqueness of the cluster-specific features in ligand space. Similarly to protein space, there are several cases where the cluster-specific interfacial features of a particular kinase cluster are indeed uniquely populated by that cluster's docked ligands. The hydrogen-bond-donor interfacial features of kinase cluster IV and the hydrogen-bond-acceptor interfacial features of kinase cluster VI, for example, are populated by ligand atoms only from the complexes of these clusters' kinase members. The remainder of the cluster-specific interfacial features in ligand space are likewise characterized by low relative probabilities of being occupied by ligands docked to kinases from other clusters.

As a validation step for the identification of unique interfacial features, we extracted features that are unique to the MAP kinases analyzed in our study. One of the identified unique hydrophobic features in protein space is the gatekeeper residue, which is not found to be a unique hydrophobic feature for the non-MAP kinase structures. This finding shows agreement with the gatekeeper hypothesis for the MAP kinases[48,49] and indicates that the feature extraction method is able to reproduce features that have been detected experimentally. In addition to the gatekeeper residue, several other unique hydrophobic features were found for the MAP kinases, suggesting that the gatekeeper does not act alone in determining binding selectivities.

Graphical Analysis of Cluster-Specific Interfacial Features. Figure 6 shows the locations of cluster-specific hydrogen bonding interfacial features in protein space for each kinase cluster. The figure also indicates the feature numbers and the quantity of hydrogen bonding interactions occurring at each feature. Cluster-specific features are distributed throughout the entire active site, and clusters have widely varying numbers of specific features and feature populations. Several unique hydrogen bonding regions are found along the center and C-terminal end of the hinge region. Figure S2 in Supporting Information contains examples of docked ligands from the Karaman et al. data set as they form hydrogen bonds at the sites of cluster-specific hydrogen bonding interfacial features. Cluster-specific hydrophobic-interaction interfacial features consist primarily of protein residues concentrated in the N-terminal lobe portion of the active site (Figure 7). In contrast, cluster-specific polar–polar-contact interfacial features comprise primarily protein side chains distributed throughout the active site (Figure S3 in Supporting Information). Protein structural elements that define the specific interfacial features of kinase clusters are listed in Table S7 in Supporting Information.

The relative frequency of the types of hydrogen bonding atoms in protein space is similar for both the set of hydrogen bonds across all eight kinase clusters and those occurring in the

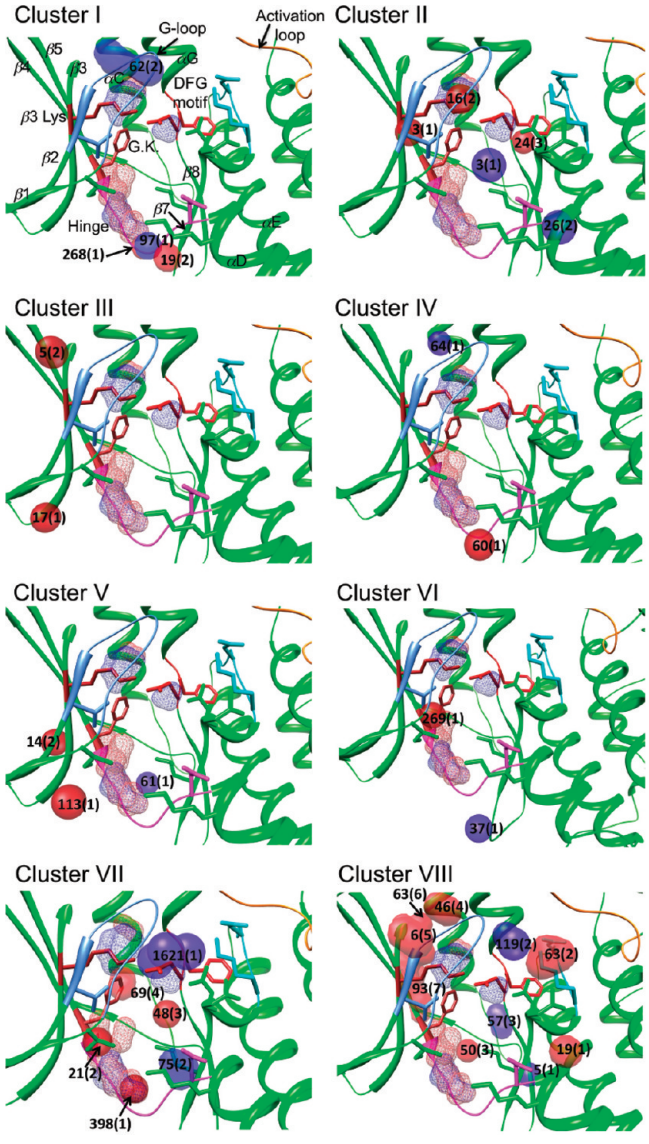


Figure 6. Cluster-specific hydrogen bonding features in protein space. Hydrogen bonding features that are unique to each kinase cluster are enclosed by solid surfaces, and global hydrogen bonding features are enclosed by mesh surfaces. Red and blue surfaces denote spaces containing high frequencies of protein hydrogen bond acceptor and donor atoms, respectively. Numbers on the panels indicate the numbers of occurrences of hydrogen bond acceptors or donors in the corresponding feature bin across all inhibitor–protein complexes in the kinase cluster. Numbers in parentheses denote feature numbers as listed in Table S7; for example, 268(1) accompanying a red surface indicates that the surface represents hydrogen bond acceptor feature 1 for the given kinase cluster and that the feature contains a total of 268 occurrences of hydrogen-bond-accepting atoms. Each panel shows protein hydrogen bonding features superimposed on the structure of human CDK2 (PDB code: 1B38) to provide a consistent reference protein structure. Structure labels in the first panel apply to all panels. G.K. signifies the gatekeeper residue.

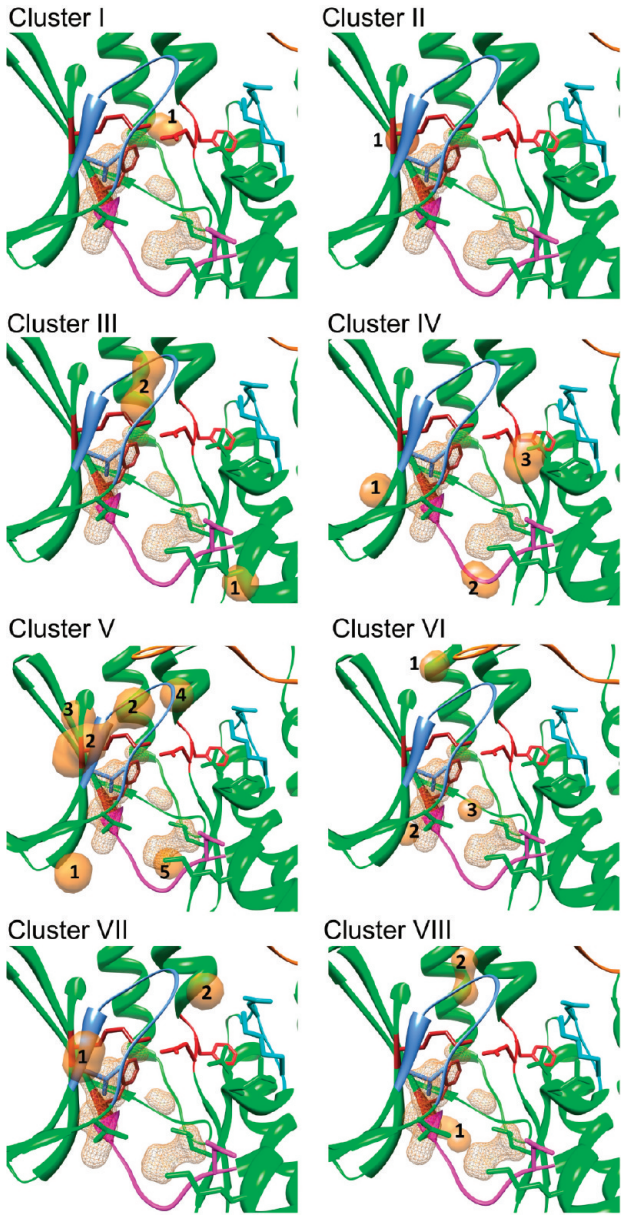


Figure 7. Cluster-specific protein features corresponding to hydrophobic interactions with docked ligands. Hydrophobic features that are unique to individual clusters are enclosed by solid surfaces; globally shared hydrophobic features are enclosed by mesh surfaces. Hydrophobic features are shown superimposed on the structure of human CDK2 (PDB code: 1B38) to provide a consistent structural reference. Numbers next to features correspond to feature numbers as listed in Table S7.

cluster-specific regions. Among all hydrogen bonds observed in the docked ligand–protein complexes, 65% of hydrogen bonding protein atoms are main-chain amide group oxygen or nitrogen atoms, while 35% of hydrogen bonding protein atoms occur in amino acid side chains. Similarly, among hydrogen bonds observed at the cluster-specific hydrogen bonding features, 71% and 29% of the bonding protein atoms are in main-chain amide groups and side chains, respectively. Furthermore, 57% of the hydrogen bonding atoms in the cluster-specific regions of protein space are hydrogen bond donors and 43% are acceptors.

Similar to interfacial features that are specific to individual kinase clusters, there exist hydrogen bonding interfacial features in ligand space that are specific to each of the four ligand clusters (Figure 8). These are spatial bins that are highly populated with ligand atoms forming intermolecular interactions with protein. Ligand cluster 1 shows a wide spatial distribution of several cluster-specific hydrogen bonding regions. The most highly populated of its cluster-specific hydrogen-bond-donating regions is situated near β3, β4, and

dx.doi.org/10.1021/jm200979x | J. Med. Chem. 2012, 55, 1926–1939

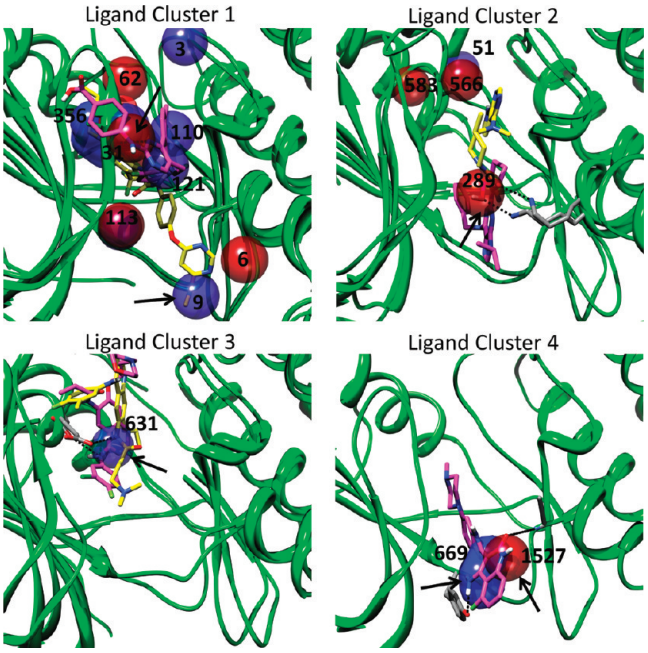


Figure 8. Hydrogen bonding features in ligand space that are specific to individual ligand clusters. Red and blue surfaces denote features that have high frequencies of hydrogen-bond-accepting and -donating atoms, respectively, in kinase inhibitor molecules across the standard set of 112 kinases. Examples of docked inhibitors belonging to the specified ligand cluster and forming hydrogen bonds with kinases are shown. Numbers on the panels denote the numbers of occurrences of hydrogen bond acceptors or donors in the corresponding feature bin across all ligand–protein complexes within the given ligand cluster. Arrows indicate hydrogen bond examples that occur at the location of cluster-specific hydrogen bonding features. Hydrogen bonds are represented by dashed lines.

β5, while the most highly populated hydrogen-bond-accepting region is located near β1 and β2 at the base of the G-loop. The features specific to ligand cluster 2 consist almost exclusively of hydrogen-bond-accepting regions, with the two most heavily populated regions positioned near the top of the G-loop. A hydrogen-bond-donating region for cluster 2 occurs adjacent to αG. Ligand cluster 3 contains a single cluster-specific hydrogen-bond-donating region situated near the C-terminus of α3. Ligand cluster 4 has a concentration of cluster-specific hydrogen-bond-donating and -accepting regions near the C-terminal end of the hinge region.

The clustering of the experimental binding data assigns to ligand cluster 4 most of the broadly binding inhibitors, including staurosporine and sunitinib. These inhibitors bind to the majority of the kinase set with moderate to high affinity (Figure 4). For ligand clusters 1–3, an average of 50% of the intermolecular hydrogen bonds occurring in the cluster-specific hydrogen bonding regions are formed with protein main-chain atoms. In contrast, 90% of the hydrogen bonds occurring in the cluster-specific regions of ligand cluster 4 are formed with main-chain atoms. Staurosporine, for example, has interactions with proteins from all eight kinase clusters that are mediated by hydrogen bonds formed almost exclusively with main-chain carbonyl oxygen atoms. This preference for main-chain atoms suggests that hydrogen bonding by inhibitors in ligand cluster 4 is less sensitive to the nature of specific side chains than is the case for ligands in other clusters. The putative reduced sensitivity to side-chain identity may underlie the broad-binding nature of the ligands in cluster 4.

Validation of Feature-Identification Procedure Using Second Inhibitor Set. Fedorov and co-workers have reported shifts in thermostabilities for 60 Ser/Thr kinase targets upon binding by 156 kinase inhibitors.[4] These shifts correlate with binding affinity, where a ΔT_m of 4 °C corresponds to K_d < 1 μM and a ΔT_m of 8 °C corresponds to K_d < 100 nM. We selected 91 kinase inhibitors from this set and docked them to our studied group of 112 kinases taken from the Karaman et al. data set. Of the 60 kinases in the data set of Fedorov et al., 28 also belong to our set of 112 analyzed kinases. Thus, we are able to compare the experimentally observed binding strengths of these shared 28 kinases with their top-scoring binding conformations as predicted by virtual docking. We compared the coincidence of type and location of predicted interactions with the cluster-specific features calculated previously for the Karaman et al. data set.

Among the simulated ligand–protein complexes of the 91 kinase inhibitors taken from the data set of Fedorov et al., those that are tight binders ($\Delta T_m \geq 4$ °C) form more intermolecular contacts with cluster-specific regions than do weak binders. Tightly binding ligands from this set form on average two hydrogen bonds and two hydrophobic contacts at the location of interfacial features that are specific for the cluster to which the interacting kinase belongs. In contrast, ligands whose binding is associated with ΔT_m < 4 °C form on average only 0.8 hydrogen bond and 1.1 hydrophobic contact with interacting kinases at the sites of cluster-specific interfacial features. Figure S4 in Supporting Information shows the conformations of three of these inhibitors bound to the tyrosine kinases ITK and LYN. The coincidence of the inhibitors' hydrogen bonding atoms with the locations of cluster-specific hydrogen bonding interfacial features is readily observed.

For the set of 91 kinase inhibitors from the study of Fedorov et al., we find that those inhibitors that form stronger interactions with a wide range of kinases also tend to form a greater share of their hydrogen bonds with main-chain atoms in their virtually docked conformations than do narrowly binding inhibitors. Among those inhibitors that have an average experimental $\Delta T_m \geq 4$ °C across the complete set of kinases in the Fedorov et al. data set, 76% of the hydrogen bonds of the docked conformations are formed with protein main-chain atoms. In contrast, among inhibitors with average experimental ΔT_m < 4 °C across the set of kinases, only 50% of modeled hydrogen bonds are formed with main-chain atoms. This observed propensity of broad binders to form hydrogen bonds with main-chain atoms is also found for the broadly binding ligands in ligand cluster 4 from the Karaman et al. data set.

Assessing Whether Cluster-Specific Features Are Also SDFs. Cluster-specific features were initially considered to be only potential SDFs. In principle, it is possible that a set of structural features is unique to a kinase cluster but does not constitute the actual determinants of binding selectivity for the associated proteins. In order to shed more light on whether the identified cluster-specific features are actual determinants of binding selectivity, we studied their relative contributions to ligand binding affinities. For the present study, this determination was limited to hydrogen bonding features that consist of protein side chains. We reasoned that for a given ligand–protein complex, removal of the cluster-specific hydrogen bonding features in protein space should have a more detrimental effect on the strength of the binding interaction

dx.doi.org/10.1021/jm200979x | J. Med. Chem. 2012, 55, 1926–1939

Table 3. Predicted and Experimentally Measured Values of $-\log(K_i)$ of Selected Compounds for Kinases CDK2, ZAP70, and PYK2[a]

	CDK2		ZAP70		PYK2	
	predicted	observed	predicted	observed	predicted	observed
aloisine	>5.0	9.28 ± 0.07	>5.0	5.58 ± 0.66	>5.0	6.03 ± 0.43
NU-6102	>5.0	8.98 ± 0.05	>5.0	<5.0	<5.0	<5.0
SC-221409	>5.0	6.97 ± 0.03	>5.0	<5.0	<5.0	<5.0
SU-11274	>5.0	6.31 ± 0.15	>5.0	<5.0	<5.0	7.98 ± 0.25
D4426	>5.0	8.28 ± 0.06	>5.0	5.93 ± 0.44	>5.0	6.23 ± 0.28
quinoxaline1	>5.0	5.67 ± 0.29	>5.0	6.08 ± 0.35	>5.0	7.39 ± 0.17
tyrphostinA23	>5.0	6.21 ± 0.38	>5.0	<5.0	<5.0	6.79 ± 0.17
scytonemin	>5.0	6.43 ± 0.20	>5.0	5.81 ± 0.42	>5.0	<5.0
dimethyladenine	>5.0	5.96 ± 0.13	>5.0	<5.0	<5.0	<5.0

[a]Compounds with $-\log(K_i) > 5.0$ (i.e. $K_i < 10\ \mu M$) were defined as binders.

than removal of all other hydrogen bonding protein side chains. In order to test this possibility, we focused on the set of complexes consisting of inhibitors from ligand cluster 4 and kinases from kinase cluster VIII, that is, cluster (VIII,4), and also on the sets of complexes from clusters (VII,4) and (II,3). Most complexes in these sets are characterized by high-affinity interactions (Figure 4C).

For the set of ligand−protein complexes in cluster (VIII,4), removing the cluster-specific hydrogen bonding side chain features by mutating the residues to Gly results in greater reductions in ligand binding affinity as predicted by the empirical scoring function XScore[62] than does removal of all other protein side chains that form hydrogen bonds with the ligand. Across the whole set of complexes taken from cluster (VIII,4), removing the cluster-specific hydrogen bonding features leads to an average increase of 650 nM in predicted K_d values, whereas eliminating all other hydrogen bonding side chains simultaneously increases predicted K_d values by an average of 70 nM. Similarly, for the set of complexes from cluster (VII,4), the corresponding increases in predicted K_d values are 370 nM and 120 nM, respectively, while for cluster (II,3) the increases are 310 nM and 50 nM. Thus, for each cluster set, eliminating intermolecular hydrogen bonds formed with protein side chains that are cluster-specific features leads to significantly greater reductions in predicted binding affinity than does removal of all other hydrogen bonding side chains. This observation suggests that the cluster-specific hydrogen bonding features contribute more to binding affinity than do the other hydrogen bonding features and so likely play a greater role in binding selectivity than do other hydrogen bonding sites. For this reason, we designate the cluster-specific features as SDFs.

Prediction of Kinase Binding for New Ligands by Machine Learning. *Predictions of Binding Affinities of Single Ligands.* We sought to assess whether the identified cluster-specific structural features of the binding interface can be used to predict the K_i of inhibitors for individual kinases. To this end, we constructed a random forest machine-learning model that uses vectors corresponding to each modeled ligand−protein complex from the Karaman et al. data set as input data. The vectors consist of the distances between all cluster-specific features for hydrogen bonding, polar−polar contacts, and hydrophobic interactions and the closest atoms in the docked ligand that participate in each corresponding interaction type. In our training set, there are 892 vectors corresponding to a 'binder' class label (experimental $K_d < 10\ \mu M$) and 2947 vectors corresponding to a 'nonbinder' class label (experimental $K_d > 10\ \mu M$). During the model training, accuracy rates for classification of binders and nonbinders are, respectively, 76% and 83%, as determined by out-of-bag (OOB) error estimation.

The binding affinity of compounds from outside the standard set of 38 kinase inhibitors for a specific kinase structure can be predicted using the trained random forest model. We selected nine such compounds and predicted the K_d values of their binding interactions with CDK2, ZAP70, and PYK2 based solely on their computationally docked poses in the active sites of these proteins. Table 3 lists the selected compounds and predictions of their binding affinity ranges, reported as $-\log(K_d) > 5.0$ or $-\log(K_d) < 5.0$ for binders and nonbinders, respectively. The random forest model predicts that 22 of the 27 ligand−kinase complexes have $-\log(K_d) > 5.0$ (binders). Only NU-6102, SC-221409, SU-11274, tyrphostinA23, and dimethyladenine are predicted to bind to PYK2 with $-\log(K_d) < 5.0$ (nonbinders).

Experimental Validation of Inhibitor K_d Predictions. We used electrospray ionization mass spectrometry (ESI-MS) to perform inhibition assays for the nine selected compounds described above against kinases CDK2, ZAP70, and PYK2 (spectra shown in Figures S5−S8 in Supporting Information). As indicated in Table 3, the experimental $-\log(K_i)$ values (K_i and K_d values interpreted the same for competitive inhibitors) of all nine compounds are >5.0 for CDK2 and hence satisfy the criterion for being classified as binders. For ZAP70 and PYK2, five and four of the compounds, respectively, had no observed inhibition ($-\log(K_i) < 5.0$). Overall, 19 of the 27 predicted K_i values were in agreement with their experimental values, yielding an accuracy rate of 70%. This accuracy rate is slightly less than the accuracy rate for predicting binders (76%) obtained from the OOB estimate during model training.

Predictions of Selectivity Profiles of Ligands and Kinases. We also evaluated how well the identified structural features can predict binding selectivity profiles of individual compounds across sets of proteins and of individual proteins across sets of compounds. To this end, we used the trained random forest model described above to predict binding interactions of 91 kinase inhibitors and 8 kinases (ERK1, ERK3, CAMK1D, CAMK2D, CDK2, CLK1, LOK, GSK3B) taken from the binding data set of Fedorov and co-workers.[4] The inhibitors are the same compounds whose docked conformations were calculated and described in an earlier section (Validation of Feature-Identification Procedure Using Second Inhibitor Set). The 728 ligand−protein combinations were predicted to correspond to binding or nonbinding interactions based on

dx.doi.org/10.1021/jm200979x | J. Med. Chem. 2012, 55, 1926−1939

having predicted K_d values <10 μM or >10 μM, respectively. For all 728 protein−ligand interaction combinations from the Fedorov data set that we analyzed, the model predicts interaction strengths with 69% accuracy, which is similar to the accuracy estimated from out-of-bag calculations during the training of the random forest model and from the experimentally validated set of nine compounds and three kinases presented above.

For predictions of binding profiles of individual kinases across the set of 91 inhibitors, there is good agreement between our predictions and the experimental values measured by Fedorov and co-workers. As shown in Table S8 in Supporting Information, there is an average of 70 ± 10% agreement between predictions and experiment for each kinase's binding profile to the set of 91 compounds. The model performs particularly well for kinases ERK1 and CAMK2D, for which it predicts binding profiles with 84% and 77% accuracy, respectively.

Similarly, predictions of binding profiles of single compounds across the set of eight kinases are in generally good agreement with experiment. For 9, 12, and 27 of the total 91 compounds, the model predicts their binding profiles to the set of eight kinases with 100%, 87.5%, and 75% accuracy, respectively, accounting for over half the analyzed compounds (Table S8). Overall, 85 of the 91 compounds have their binding profiles estimated with ≥50% accuracy.

In addition, we calculated whether the accuracy of compound binding profile prediction is greater than what would be expected simply on the basis of independent event probabilities. Figure S9 in Supporting Information shows cumulative probabilities of having given numbers of total correct predictions (out of eight) for individual compound binding profiles to the eight kinases. As shown in the figure, the probability curve that corresponds to the observed prediction accuracies is shifted to the right of the curve that would be expected if the binding profiles were predicted with an accuracy simply matching independent event probability. For instance, when the probability of success for a single trial matches the overall prediction accuracy for the trained model, the expected binomial distribution probability of having ≥6 correct predictions for a given compound's binding profile is 0.32. In fact, the observed probability of having ≥6 correct predictions is 0.53. Similarly, the expected and observed probabilities for having ≥7 correct predictions are 0.11 and 0.23, respectively, and those for having eight correct predictions are respectively 0.02 and 0.10. In all cases for cumulative probabilities of having at least 4 or more correct predictions out of the 8 predictions that constitute a compound's binding profile, the model performs with scaling that is better than that of independent probability.

In similar fashion, the accuracy of prediction of the binding panels of the kinases to the 91 compounds also exceeds that which would be expected based on independent event probability. For four of the eight kinases, the binomial distribution probability of achieving the observed accuracy is <0.01, and for only two of the kinases is the probability >0.5.

Taken together, these results show that the observed accuracy of the model in predicting binding profiles is notably greater than what would be expected if profiles were forecast following simple independent event probability, that is, as though the pattern of activity of a compound or protein comprised separate, unrelated individual predictions. This improvement over binomial distribution behavior suggests that the identified SDFs indeed play a role in regulating binding selectivity and that the random forest model successfully incorporates these features.

Predictions of Cluster Assignments. The vectors used in conjunction with the random forest model to predict binding affinities were likewise used as input for a random forest model to predict ligand cluster assignments for the ligands of all virtually docked ligand−protein complexes. The model was trained using known ligand cluster numbers (1−4) as class descriptors. During training, the OOB accuracy rate for cluster classification is 88%. Similarly, the classification accuracy as determined by leave-one-out analysis across the whole set of modeled kinases is 89%. We also constructed a random forest model using the same input vectors and kinase cluster numbers (I−VIII) as class descriptors in order to predict kinase cluster assignments. The classification accuracy associated with this model as determined by both OOB during model training and by leave-one-out analysis is 99%. Overall, 88% of ligand−protein complexes are assigned to both the correct ligand cluster corresponding to the ligand and the correct kinase cluster corresponding to the protein. This result indicates that, for a ligand−protein pair, the cluster assignment of both the inhibitor and the kinase can be predicted with considerable accuracy based solely on the docked conformation of the inhibitor relative to the cluster-specific interface binding features.

Further, given the correct cluster assignment for a ligand−protein complex, the relative binding affinity of the ligand is suggested by the average affinities of other members of the cluster. For example, the average experimental K_d values of complexes assigned by the random forest model to clusters (VII,1), (VII,2), (VII,3), and (VII,4) during the leave-one-out procedure are 0.1 μM, 3 μM, 2 μM, and 0.06 μM, respectively. The average experimental K_d values for all the members of these same clusters are 2.6 μM, 9.2 μM, 5 μM, and 0.5 μM, respectively. Similarly, the average experimental K_d values for complexes assigned by the model to clusters (II,1), (II,2), (II,3), and (II,4) are 5 μM, 3 μM, 0.03 μM, and 5 μM, while the average experimental K_d values for all members of these clusters are 7.2 μM, 7.4 μM, 0.05 μM, and 7.5 μM. In both cases, we observe consistency between the trends for the average experimental binding affinities across all members of a cluster and the average experimental binding affinities of complexes assigned to the cluster by the random forest model.

■ DISCUSSION

Our findings bolster our hypothesis that large groups of proteins and ligands can be clustered based on their binding profile similarities and that these clusters, in conjunction with ligand docking, can be used to extract features of the binding interface between proteins and ligands that underlie similarities and differences in binding profiles. As we demonstrate using kinases as a model, clustering of proteins and ligands according to similarity of experimental K_d values proves useful as a first step toward identifying these features. In conjunction with binding-affinity heat maps, clustering reveals clear groupings of ligand−protein binding patterns. For example, nearly all broadly binding inhibitors, such as staurosporine and sunitinib, cluster together in ligand space and bind to all protein clusters, while other inhibitors tend to bind only to specific protein clusters. For each protein cluster, we used in silico docking to generate ligand−protein complex structures and identified spatial locations, referred to as interfacial features, in the

binding interface that participate with high frequency in intermolecular interactions, including hydrogen bonding, polar−polar contacts, and hydrophobic interactions. A large number of such interfacial features are shared globally for kinases across all eight kinase clusters. Conversely, many features are found to be specific to kinases belonging to a single cluster. These SDFs drive the cluster assignments of the proteins and ligands and demonstrate that there are numerous chemical features in both proteins and ligands that govern binding selectivity. Interestingly, the presence of these multiple features indicates that selectivity and cluster assignment are determined not by single structural features, such as the identity of the gatekeeper residue in the MAP kinases, but by combinations of features.

Cluster-specific hydrogen bonding interfacial features consisting of protein side chains were analyzed in terms of their contributions to ligand binding affinity relative to other hydrogen bonding side chains. For each of three sets of complexes that correspond to high-affinity groupings in Figure 4C, results of energy calculations suggest that the cluster-specific hydrogen bonding features make a significantly greater contribution to binding affinity than do the other hydrogen bonding side chains. This observation implies that the cluster-specific features are, at least in terms of hydrogen bonding interactions, determinants of selectivity and led us to designate the cluster-specific features as SDFs.

Identification of SDFs in the binding interface that are specific to individual groups of kinases provides an easily visualized and chemically intuitive model for the basis of ligand binding selectivity. The locations of SDFs in kinases from a given kinase cluster indicate the amino acid residues (when examining SDFs in protein space) or ligand functional groups (when examining SDFs in ligand space) that contribute to each type of interaction with the kinases. Knowledge of these residues or functional groups can potentially be applied to design small molecules that selectively target the kinases from that cluster.

SDFs can also be used to model the strength of binding interactions. Machine-learning models that employ identified SDFs as inputs are predictive of the binding affinity of compounds to specific kinases. A random forest model that uses binding conformations of docked compounds relative to the cluster-specific interfacial features predicts with good accuracy whether an individual ligand−protein interaction is characterized by K_i < 10 μM. Applying this model, we can also predict reasonably well the selectivity of a single compound across a set of kinases or, alternatively, the selectivity of a single kinase across a set of compounds, as demonstrated for a collection of 91 kinase inhibitors and 8 kinases taken from the binding data set of Fedorov and co-workers. However, despite the ability of the random forest model to predict binding profiles with decent accuracy, there is room for improvement. Current work in our laboratory focuses on additional descriptors to add to the descriptor vectors that are used in conjunction with the random forest model in order to improve further the accuracy of binding profile prediction.

The presented methodology represents a shift from certain other recently reported QSAR models for kinase binding. For instance, Sheridan and co-workers have developed a model[63] that accurately predicts the overall similarity in the binding profiles of kinase pairs for given sets of ligands, but the model does not predict binding behavior for individual ligands. To our knowledge, the present study is the first to demonstrate accurate prediction of kinase binding affinities using 'selectivity filters' in both protein space and ligand space that have been computationally derived from a large set of virtually docked ligand−protein complexes. The use of derived selectivity filters to predict ligand binding affinities provides an alternative approach to the use of energy force fields and statistical scoring functions.

Our methodology is based on several assumptions. First, it assumes that the clustering protocol, which is based solely on similarities of binding affinity profiles, also reflects structural and chemical differences between kinase active sites. Second, as the approach depends on identifying specific structural features of the ligand−protein interfaces of kinases, it necessarily assumes that the important forces for determining binding interactions are attributable to the residues immediately surrounding the active site. Third, it presupposes accurate in silico docking poses for ligand−protein complexes. Nonetheless, the relatively high accuracy of our docking protocol in self-docking tests leaves us confident that a sizable majority of the predicted docking poses used for identifying interface features are likewise accurate. Moreover, we find that the clustering method indeed captures notable differences between the active sites of kinases belonging to different kinase clusters in terms of structure and chemical properties, as indicated by both the PM*Score* algorithm[53] and ClustalW[54−56] amino acid sequence alignment.

Whereas previous studies of SDFs have focused on the structures and microenvironments of kinases themselves, our investigation takes a broader view and explores features of the ligand−protein binding interface at both the protein structural level and the ligand structural level. The predominant features that are pinpointed by the present methodology are consistent with those that are found by examining large numbers of individual kinase structures on a case-by-case basis, as analyzed elsewhere.[52] Given that our methodology is applicable to other protein families, this finding shows that the detection of binding-interface features of a large set of proteins, including whole enzyme classes and protein families, can be automated and need not necessarily depend on manual inspection of numerous individual experimental structures.

Current work in our laboratory is focused on utilizing the identified structural features specific to individual kinase clusters for specific applications. The set of cluster-specific ligand−protein interface features may be conceived of as hot-spots within the protein active site, and molecular fragments placed at the hot-spot locations can be linked to generate complete molecules.[64−67] This de novo fragment-based drug discovery approach may yield novel selective kinase inhibitors and aid in toxicity prediction. In addition, we are screening the current release of the Protein Data Bank for protein structures, including nonkinases, containing active sites whose main chain and side chains are arranged such that they can form atomic contacts with the determined interfacial feature locations. In principle, this screening may identify proteins whose activity can be modulated by kinase inhibitors belonging to specific ligand clusters in our analyzed data set. Unexpected cross-reactivity of approved inhibitors with pharmacologically interesting targets represents the potential to repurpose existing drugs for alternative therapeutic needs.

■ ASSOCIATED CONTENT

Supporting Information

Computational and experimental methods. List of kinases with available PDB structures used for virtual docking. List of kinases belonging to kinase clusters I−VIII. List of ligands belonging to ligand clusters 1−4. List of ligand−kinase complexes used to assess docking accuracy. Protein structural elements that define SDFs. Structural similarity of kinase binding sites. Examples of docked kinase inhibitors forming hydrogen bonds. Locations of cluster-specific polar−polar contact features. Docked conformations of ligands from alternative data set. Mass spectral data analysis results and mass spectra for enzyme activity assays. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jsschoe@sandia.gov. Tel: 925-294-2955.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge support from the Defense Threat Reduction Agency CB Basic Research Program and the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

■ ABBREVIATIONS USED

SDF, specificity-determining feature

■ REFERENCES

(1) Pastor, M.; Cruciani, G. A novel strategy for improving ligand selectivity in receptor-based drug design. *J. Med. Chem.* **1995**, *38*, 4637−4647.

(2) Melnick, J. S.; Janes, J.; Kim, S.; Chang, J. Y.; Sipes, D. G.; Gunderson, D.; James, L.; Matzen, J. T.; Garcia, M. E.; Hood, T. L.; Beigi, R.; Xia, G.; Harig, R. A.; Asatryan, H.; Yan, S. F.; Zhou, Y.; Gu, X.-J.; Saadat, A.; Zhou, V.; King, F. J.; Shaw, C. M.; Su, A. I.; Downs, R.; Gray, N. S.; Schultz, P. G.; Warmuth, M.; Caldwell, J. S. An efficient rapid system for profiling the cellular activities of molecular libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3153−3158.

(3) Bain, J.; Plater, L.; Elliott, M.; Shpiro, N.; Hastie, C. J.; McLauchlan, H.; Klevernic, I.; Arthur, J. S. C.; Alessi, D. R.; Cohen, P. The selectivity of protein kinase inhibitors: a further update. *Biochem. J.* **2007**, *408*, 297−315.

(4) Fedorov, O.; Marsden, B.; Pogacic, V.; Rellos, P.; Müller, S.; Bullock, A. N.; Schwaller, J.; Sundström, M.; Knapp, S. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20523−20528.

(5) Bamborough, P.; Drewry, D.; Harper, G.; Smith, G. K.; Schneider, K. Assessment of chemical coverage of kinome space and its implications for kinase drug discovery. *J. Med. Chem.* **2008**, *51*, 7898−7914.

(6) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(7) Rhee, S.-Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298−303.

(8) Rawlings, N. D.; Barrett, A. J.; Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **2010**, *38*, D227−D233.

(9) Senger, S.; Leach, A. R. SAR Knowledge Bases in Drug Discovery. *Annu. Rep. Comput. Chem.* **2008**, *4*, 203−216.

(10) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(11) Santos-Filho, O. A.; Cherkasov, A. Using molecular docking, 3D-QSAR, and cluster analysis for screening structurally diverse data sets of pharmacological interest. *J. Chem. Inf. Model.* **2008**, *48*, 2054−2065.

(12) Huang, D.; Zhou, T.; Lafleur, K.; Nevado, C.; Caflisch, A. Kinase selectivity potential for inhibitors targeting the ATP bindng site: a network analysis. *Bioinformatics* **2010**, *26*, 198−204.

(13) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(14) Fliri, A. F.; Loging, W. T.; Volkmann, R. A. Analysis of system structure-function relationships. *ChemMedChem* **2007**, *2*, 1774−1782.

(15) Edwards, A. Large-scale structural biology of the human proteome. *Annu. Rev. Biochem.* **2009**, *78*, 541−568.

(16) Terwilliger, T. C.; Stuart, D.; Yokoyama, S. Lessons from structural genomics. *Annu. Rev. Biophys.* **2009**, *38*, 371−383.

(17) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180−190.

(18) Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. S. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1544*, 350−357.

(19) Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteochemometric modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.* **2002**, *61*, 1465−1475.

(20) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics--structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243−357.

(21) Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. S. Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. *Bioinformatics* **2005**, *21*, 4289−4296.

(22) Ortiz, A. R.; Gomez-Puertas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Viñas, E.; Morreale, A.; Murcia, M.; Wang, K. Computational approaches to model ligand selectivity in drug design. *Curr. Top. Med. Chem.* **2006**, *6*, 41−55.

(23) Kontijevskis, A.; Wikberg, J. E. S. Computational proteomics analysis of HIV-1 protease interactome. *Proteins* **2007**, *68*, 305−312.

(24) Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. S. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinf.* **2008**, *9*, 181.

(25) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767−779.

(26) Kontijevskis, A.; Petrovska, R.; Yahorava, S.; Komorowski, J.; Wikberg, J. E. S. Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates. *Bioorg. Med. Chem.* **2009**, *17*, 5229−5237.

(27) Weill, N.; Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model* **2009**, *49*, 1049−1062.

(28) Hu, Y.; Bajorath, J. Exploring target-selectivity patterns of molecular scaffolds. *ACS Med. Chem. Lett.* **2010**, *1*, 54−58.
(29) Strömbergsson, H.; Lapinsh, M.; Kleywegt, G. J.; Wikberg, J. E. S. Towards proteome-wide interaction models using the proteochemometrics approach. *Mol. Inf.* **2010**, *29*, 499−508.
(30) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.
(31) Pettit, F. K.; Bowie, J. U. Protein surface roughness and small molecular binding sites. *J. Mol. Biol.* **1999**, *285*, 1377−1382.
(32) DeLano, W. L. Unraveling hot spots in binding interfaces: Progress and challenges. *Curr. Opin. Struct. Biol.* **2002**, *12*, 14−20.
(33) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518−2525.
(34) Fabbro, D.; Garcia-Echeverria, C. G. Targeting protein kinases in cancer therapy. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 701−712.
(35) Fabbro, D.; Ruetz, S.; Buchdunger, E.; Cowan-Jacob, S. W.; Fendrich, G.; Liebetanz, J.; Mestan, J.; O'Reilly, T.; Traxler, P.; Chaudhuri, B.; Fretz, H.; Zimmermann, J.; Meyer, T.; Caravatti, G.; Furet, P.; Manley, P. W. Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. *Pharmacol. Ther.* **2002**, *93*, 79−98.
(36) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28−39.
(37) Myers, M. R.; He, W.; Hulme, C. Inhibitors of tyrosine kinases involved in inflammation and autoimmune diseases. *Curr. Pharm. Des.* **1997**, *3*, 473−502.
(38) Cohen, P. The development and therapeutic potential of protein kinase inhibitors. *Curr. Opin. Chem. Biol.* **1999**, *3*, 459−465.
(39) Asadullah, K.; Volk, H.-D.; Sterry, W. Novel immunotherapies for psoriasis. *Trends Immunol.* **2002**, *23*, 47−53.
(40) Orchard, S. Kinases as targets: prospects for chronic therapy. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 713−717.
(41) Kumar, S.; Boehm, J.; Lee, J. C. p38 MAP kinases: key signalling molecules as therapeutic targets for inflammatory diseases. *Nat. Rev. Drug Discovery* **2003**, *2*, 717−726.
(42) Saklatvala, J. The p38 MAP kinase pathway as a therapeutic target in inflammatory disease. *Curr. Opin. Pharmacol.* **2004**, *4*, 372−377.
(43) Gabriele, A.; King, G. L. Protein kinase c inhibitors in the treatment and prevention of diabetic complications. *Curr. Opin. Endocrinol. Diabetes* **2001**, *8*, 197−204.
(44) Kikuchi, Y.; Yamada, M.; Imakiire, T.; Kushiyama, T.; Higashi, K.; Hyodo, N.; Yamamoto, K.; Oda, T.; Suzuki, S.; Miura, S. A Rho-kinase inhibitor, fasudil, prevents development of diabetes and nephropathy in insulin-resistant diabetic rats. *J. Endocrinol.* **2007**, *192*, 595−603.
(45) Weinblatt, M. E.; Kavanaugh, A.; Burgos-Vargas, R.; Dikranian, A. H.; Medrano-Ramirez, G.; Morales-Torres, J. L.; Murphy, F. T.; Musser, T. K.; Straniero, N.; Vicente-Gonzales, A. V.; Grossbard, E. Treatment of rheumatoid arthritis with a syk kinase inhibitor: A twelve-week, randomized, placebo-controlled trial. *Arthritis Rheum.* **2008**, *58*, 3309−3318.
(46) Mazanetz, M. P.; Fischer, P. M. Untangling tau hyperphosphorylation in drug design for neurodegenerative diseases. *Nat. Rev. Drug Discovery* **2007**, *6*, 464−479.
(47) Harmon, B.; Campbell, N.; Ratner, L. Role of Abl kinase and the Wave2 signaling complex in HIV-1 entry at a post-hemifusion step. *PLoS Pathog.* **2010**, *6*, e1000956.
(48) Gum, R. J.; McLaughlin, M. M.; Kumar, S.; Wang, Z.; Bower, M. J.; Lee, J. C.; Adams, J. L.; Livi, G. P.; Goldsmith, E. J.; Young, P. R. Acquisition of sensitivity of stress-activated protein kinases to the p38 inhibitor, SB 203580, by alteration of one or more amino acids within the ATP binding pocket. *J. Biol. Chem.* **1998**, *273*, 15605−15610.
(49) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Beguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity of MAP kinases. *Structure* **1998**, *6*, 1117−1128.
(50) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.
(51) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.
(52) Liao, J. J. L. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J. Med. Chem.* **2007**, *50*, 409−424.
(53) Yeturu, K.; Chandra, N. PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinf.* **2008**, *9*, 543−559.
(54) Thompson, J. D.; Higgins, D. G.; Gibson, T. J.; Clustal, W Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.
(55) Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31*, 3497−3500.
(56) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947-2948.
(57) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.
(58) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(59) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.
(60) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293−2302.
(61) May, A.; Zacharias, M. Protein−ligand docking accounting for receptor side chain and global flexibility in normal modes: Evaluation on kinase inhibitor cross docking. *J. Med. Chem.* **2008**, *51*, 3499−3506.
(62) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(63) Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. QSAR models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *J. Chem. Inf. Model.* **2009**, *49*, 1974−1985.
(64) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660−672.
(65) Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: leads by design. *Drug Discovery Today* **2005**, *10*, 987−992.
(66) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(67) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211−219.