Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe

Francesca Milletti[†] and Anna Vulpetti*

CADD, Global Discovery Chemistry, Novartis Institutes for Biomedical Research, CH4002 Basel, Switzerland

Received April 3, 2010

Polypharmacology is receiving increasing attention in the pharmaceutical industry, since finding new targets of a compound is useful not only for anticipating possible side effects but also for opening new therapeutic opportunities. Thus, while system biology and personalized medicine are becoming increasingly important, there is an urgent need to map the inhibition profile of a compound on a large panel of targets by using both experimental and computational methods. This is especially important for kinase inhibitors, given the high similarity at the binding site level for the 518 kinases in the human genome. In this paper, we propose and validate a new method to predict the inhibition map of a compound by comparison of binding pockets. We used a subset of the Ambit panel for the validation—17 inhibitors with K_d measured on 189 kinases—and found that on average 37% of kinases inhibited with $K_d < 10\ \mu$M were retrieved at 10% ROC enrichment. These results make this method particularly suitable to rationalize and optimize the selectivity profile of a compound. In addition, the method was extended to explore all the proteins in the PDB by using as queries pockets occupied by compounds of biological interest (ATP and various marketed drugs). The profiling of compounds against the protein universe revealed that striking structural similarities at the subpocket level (RMSD < 0.5 Å) may also occur among targets with different folds, which can be exploited not only to predict off-target effects but also to design novel inhibitors for the target of interest.

INTRODUCTION

The classical view of one-target one-drug is being challenged today by increasing evidence that a drug can hit more than one target, which may open new therapeutic opportunities.[1] For example, only after Imatinib had entered clinical trials it was discovered that it inhibits Kit in addition to the already known Abl1 and Pdgf receptors. This prompted its use in the treatment of gastrointestinal stromal tumors.[2] However, while compounds that hit more than one target could be advantageous, highly promiscuous compounds are generally undesirable because of their high toxicity. For this reason, efforts are directed at developing compounds that hit selectively only specific targets, a task that is particularly challenging for protein families such as kinases, proteases, and ligases because the binding sites of members of the same family are highly similar.

These trends, combined with the large and increasing number of 3D structures in the Protein Data Bank[3] (nearly 63 000), have drawn much attention to methods that classify proteins by analyzing their binding site. A binding-site-based classification of proteins is supposed to improve traditional sequence-based classifications, as shown by many findings of proteins that have low overall sequence similarity, but high similarity at the binding site. Furthermore, the analysis of the 3D arrangement of atoms in the binding site is particularly useful for taking into account conformational changes, which cannot be captured by sequence-based methods.

Various approaches for comparing protein pockets have been reported in recent years.[4] FLAP[5] describes protein pockets by using four-point pharmacophoric descriptors based on GRID[6] molecular interaction fields. Cavbase[7] and SiteEngine[8,9] use three-point pharmacophoric descriptors based on pseudocenters that encode acceptor, donor, and hydrophobic properties of the atoms in the protein. SuMo[10] uses three-point pharmacophoric descriptors like Cavbase and SiteEngine, but these descriptors are based on more specific atom types. The approach by Kinnings and Jackson[11] is based on a geometric hashing method that first identifies matching atoms between pairs of targets and then calculates a similarity score based on the relative number of matches. Another method, IsoCleft,[12] uses an efficient graph-matching-based algorithm to detect 3D atomic similarities. The approach described by Konrat[13] is based on a meta-structural similarity (conservation) in the protein−ligand interaction sites of proteins. The meta-structure description requires only sequence information and is defined as an intricate network of interacting residues organized as a multispherical entity. These methods were validated by using different criteria, and therefore it is very difficult to identify directions for an optimal approach, but interesting examples of high 3D similarity between proteins with low sequence identity have been reported, such as in the case of Pdk1 and Syk, which are kinases that have only 13% sequence identity,[11] and the case of cyclooxygenase cox-2 and the carbonic anydrase (CA-II), which are totally unrelated targets showing nanomolar affinity toward Celecobix and Vadecoxib.[14]

It is important to recognize the limits of binding-site-based classifications to predict new targets of a compound. First, a ligand may change conformation upon binding to different

* Corresponding author e-mail: anna.vulpetti@novartis.com.

[†] Current address: Hoffmann-La Roche, 340 Kingsland St., Nutley, New Jersey.

10.1021/ci1001263
Published on Web 07/28/2010

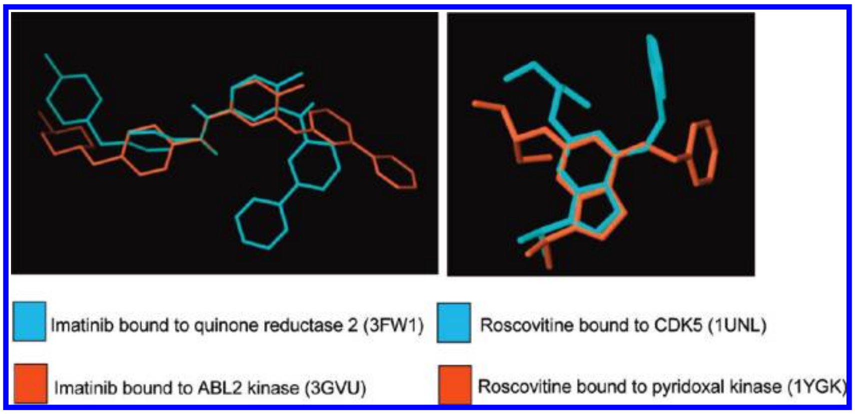

🟦 Imatinib bound to quinone reductase 2 (3FW1)
🟥 Imatinib bound to ABL2 kinase (3GVU)
🟦 Roscovitine bound to CDK5 (1UNL)
🟥 Roscovitine bound to pyridoxal kinase (1YGK)

Figure 1. Different ligand conformations for Imatinib and Roscovitine bound to kinase and nonkinase targets.

targets, and therefore related targets can have significantly different binding pockets. For example, it was shown that dual Src and PI3K pyrazolopyrimidine inhibitors bind to these two targets with a rotation of about 90° of the R-aryl side chain.[15] This effect is definitively more striking for targets of different families. For example, Figure 1 shows the different conformation of the crystallographic 3D structures of Imatinib bound to Abl2 and to Quinone reductase 2 and of Roscovitine bound to Cdk5 and to pyridoxal kinase.

Second, higher energy protein conformations may cause unfavorable binding that cannot be predicted by the 3D structure alone. A well-documented case is that of Imatinib bound to Src kinase. Although a PDB structure of Imatinib bound to Src is available (2OIQ), and the corresponding binding pocket is virtually identical to that of Abl and Kit, Imatinib has nanomolar affinity for Abl and Kit, but micromolar affinity for Src. This difference has been explained by the energy penalty paid by Src, compared to Abl and Kit, to adopt the DFG-out conformation required for binding.[16−18]

Third, in the PDB, there are multiple protein structures for the same target, but an exhaustive sampling is generally not available. In the case of kinases, this can be a problem if the structure available is in an active state, but the ligand binds to the inactive state.

Last, flexible side chains in the protein pocket may add noise in the description of the cavity, and they should be treated carefully for binding pocket comparison.

While it is important to recognize the limitations described above, it is also important to stress that in the pharmaceutical industry there is a strong interest in targets once regarded as not tractable for selectivity reasons, such as kinases. Therefore, methods focused on the optimization of the selectivity profile of compounds within a family are highly desired. Furthermore, screening a pocket against the whole PDB opens the possibility of discovering unexpected similarities of targets with different folds and thus using this information for the design of novel inhibitors.

In this paper, we rationalize the inhibition map of kinase inhibitors from the Ambit panel[19] by using a novel method that compares protein pockets. The Ambit data constitute the most comprehensive public study of kinase inhibitor selectivity to date and reveal a wide diversity of interaction patterns: K_d values for 38 kinase inhibitors across a panel of 317 kinases representing >50% of the predicted human protein kinome are reported. The paper is organized as follows: first, the developed method to calculate protein pocket similarity and to align the pockets is described. The pocket similarity method requires both the 3D structure of the target and that of the ligand in the complex, in order to consider only the protein atoms involved in the interaction. For that reason, the validation of this method is reported using a subset of the Ambit panel—17 inhibitors with K_d data for 189 different kinases—those for which at least one crystal structure is available. On average, 37% of the kinases with $K_d < 10\ \mu M$ were retrieved at 10% ROC enrichment, and it was possible to identify the most promiscuous compounds. Second, the results obtained by using the developed target-based approach are compared to those obtained from ligand-based methods, and advantages and disadvantages of the two approaches are discussed. Last, the search of structurally related pockets is extended to the whole PDB by using as queries not only kinase inhibitors but also ATP and 10 compounds of biological interest. The discovery of proteins with different folds but nearly identical binding sites is reported.

METHODS

1. New Method to Predict Polypharmacology by Pocket Similarity. The general approach for comparing pockets was inspired from a method developed by Belongie et al.[20] for shape recognition. The basic idea behind the method by Belongie et al. is to pick n points on the contour of a shape (Figure 2a) and, for each point, to generate a descriptor that encodes its "shape context" by counting the occurrences of points within spheres of radius r (Figure 2b,c). The "shape context" descriptor describes the coarse distribution of the rest of the shape with respect to a given point of the shape. Finding correspondence between the two shapes is equivalent to finding a pair of points in the two shapes having the same "shape context", i.e., the best match. This approach was used to find similar shapes on the MNIST data set of handwritten digits[21] and yielded an error rate of 0.63%, which was the lowest at the time of publication.

The method proposed by Belongie et al. has inspired the development of an approach for the 3D alignment of small molecules by using a list of atom equivalences obtained by "shape context"-based descriptors centered on each atom of a molecule.[22] This approach for the alignment of molecules, which was developed by Richmond et al.,[22] uses the list of atom equivalences and the Procrustes method[23] to obtain the molecular alignment. The method presented in this paper builds on the experience of Belongie et al.'s for the cal-

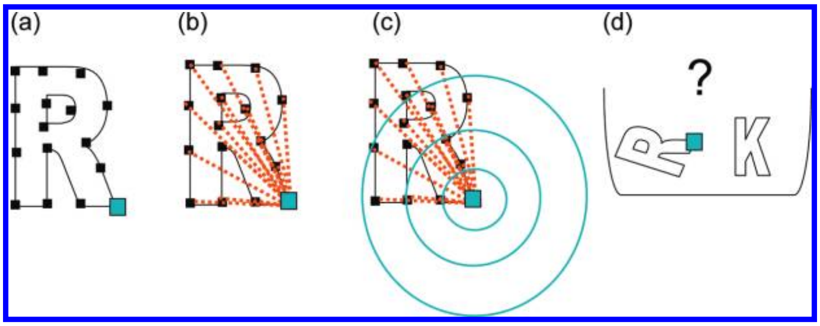


Figure 2. Shape-context-based descriptors as defined in the method by Belongie et al.

Table 1. Atom Types Used for Describing the Pocket

atom type	primary
C	Arg(CB, CD, CG) Met(CB), Phe(CB), Leu (CB, CD1, CD2M CG), Trp(CB), Asp(CB), Lys(CB,CE), His(CB), Val(CB,CG1,CG2), Gln(CB,CD,CG), Ala(CB), Glu(CB), Pro(CB,CD,CG), Cys(CB), Tyr(CB), Asn(CB), Ile(CB,CD1,CG1,CG2)
Car	Phe(CD1,CD2,CE1,CE2,CG,CZ), Trp(CD1, CD2, CE2, CE3, CG, CH2, CZ2M CZ3), His(CD2,CE1), Tyr (CD1,CD2, CE1,CE2,CG,CZ)
Carg	Arg(CZ, CE, CG)
N	all amidic N, Gln(), Asn()
ND1	His(ND1)
NE2	His(NE2)
Nlys	Lys(NZ)
Ntrp	Trp(NE1)
O	all amidic carbonyl
Ocoo	Glu(CG)
Ooh	Ser(CB), Thr(CB)
Otyr	Tyr(OH)
S	Cys(SG)

atom type	secondary
Ccoo	Asp(CB), Glu(CG)
Cgln	Gln(CG), Asn(CB)
Hyd	Arg(CZ), Met(SD), Phe(CG,CZ), Leu(CG), Trp(CE3,CG,CZ2), His(CG), Val(CB), Pro(CG), Cys(SG), Tyr(CG,CZ), Ile(CB)

culation of pocket similarity and of Richmond et al.'s for the calculation of the pocket alignment.

The general problem of predicting new targets of a compound from a classification scheme that clusters targets by their similarity at the binding site level is complex, because the inhibition map of different compounds changes depending on the residues directly involved in the binding. For example, by analyzing the Ambit data set, which contains inhibition data of 38 compounds tested on 317 kinases, we conclude that the portion of the ATP pocket (subpocket) occupied by Staurosporine in ErbB4 is highly similar to that of almost all kinases (Staurosporine hits 288 kinases out of 317 with a $K_d < 10\ \mu$M), whereas the subpocket occupied by Lapatinib in ErbB4 is similar only to the subpockets in ErbB2 and Egfr. Therefore, it is important to select the appropriate subpocket when comparing binding sites to predict the actual targets of a compound. Because the main aim of this work is not protein classification, but prediction and rationalization of the inhibition map of a compound, the basic requirement of the method is the availability of a crystal structure of the target in complex with the inhibitor of interest.

The computational procedure for predicting the inhibition map of a compound can be divided into two fundamental steps: database creation and pocket screening. Database creation requires first the generation of a database of protein pockets and then the calculation of fingerprints that describe each pocket. The second step—pocket screening—requires first the generation of a fingerprint that encodes the query pocket, which is the portion of the pocket that contains the inhibitor considered. Then, a raw similarity score is calculated between the query pocket and all the pockets in the database. Pockets are subsequently aligned to the query pocket to visualize the regions that are the most similar, and the RMSD obtained from the alignment is used to correct the raw similarity score.

1.1. Creation of a Database of Protein Pockets. Pocket Detection. Protein pockets are detected by using the program FlapSite.[24] FlapSite works by determining the degree of buriedness of points around the protein surface to identify clusters of points that belong to a real pocket. It also calculates GRID molecular interaction fields using the "DRY" probe to bias the geometric approach toward hydrophobic regions of the protein. The pocket detected by FlapSite, which is represented by dummy atoms in the interior of the pocket (2 Å from surface atoms), is converted into a PDB file that contains the protein atoms at less than 3 Å from the surface of the pocket detected by FlapSite.

Atom Type Conversion. The residues and atom types of the protein in the pocket are converted into atom types that define their physical chemical properties, as reported in Table 1. To encode similarities between hydrophobic atoms that otherwise would be represented as distinct ("Car", "C", and "S"), a secondary atom type, "Hyd", which is common to all hydrophobic groups, is added. As shown in Table 1, the atom type "Hyd" is activated only in specific atoms of a hydrophobic residue. For example, the "Hyd" atom type is active only at the CB of Val and at CZ and CG of Phe. The secondary atom types are always activated in addition to the primary atom type. Hydrogen atoms were not included among the descriptors because their position is not available from the PDB in most cases. Ionizable groups were characterized by separate atom types; for example, Nlys was used for NZ of Lys and Carg for CZ of Arg. All water molecules in the PDB structures were disregarded, given the difficulty in separating conserved waters from labile waters in a large scale study that involves PDB structures in the apo form and in complex with different ligands.

Side Chain Flexibility. Because side chain flexibility may affect the similarity score between two pockets without being determinant for binding, the physical chemical features of some residues (Ser, Thr, Asp, Gln) are encoded in one of the atoms closer to the carbon α (see Ocoo, Ooh, Ccoo, and Cgln of Table 1). For example, Figure 3 shows that the Ooh

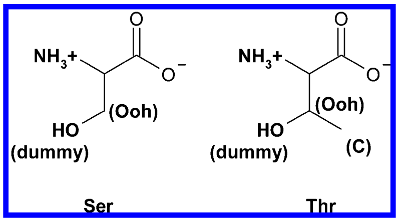
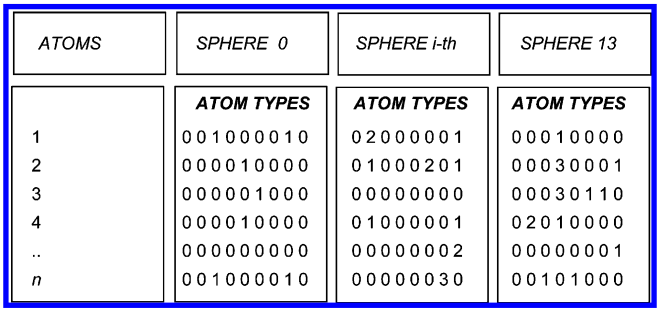


Figure 3. Dummy atom types are assigned to atoms in flexible side chains to disregard structural differences caused by side chain flexibility.



ATOMS	*SPHERE 0*	*SPHERE i-th*	*SPHERE 13*
	ATOM TYPES	***ATOM TYPES***	***ATOM TYPES***
1	0 0 1 0 0 0 0 1 0	0 2 0 0 0 0 0 1	0 0 0 1 0 0 0 0
2	0 0 0 0 1 0 0 0 0	0 1 0 0 0 2 0 1	0 0 0 3 0 0 0 1
3	0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0	0 0 0 3 0 1 1 0
4	0 0 0 0 1 0 0 0 0	0 1 0 0 0 0 0 1	0 2 0 1 0 0 0 0
..	0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 2	0 0 0 0 0 0 0 1
n	0 0 1 0 0 0 0 1 0	0 0 0 0 0 0 3 0	0 0 1 0 1 0 0 0

Figure 4. Fingerprint to describe binding pockets.

atom type of Ser and Thr is encoded in the CB atom, whereas the actual OH atoms are represented by dummy atoms and are not explicitly used in the fingerprint description of the binding site. This approach was useful for reducing the noise by removing the contribution of different orientations of these side chains.

Fingerprint. To describe the whole protein pocket, n fingerprints are generated for each of the n atoms lining the cavity. Each fingerprint encodes the occurrence of the different atom types within concentric spherical layers of increasing radius centered at each of the n atoms. Each sphere has a radius r_i based on the formula

$$r_i = r_{i-1} + \frac{(i+3)^2}{100} \tag{1}$$

where i ranges from 1 to 13. The radius of the first sphere (r_0) is 2 Å, and that of the last sphere (r_{13}) is 16.8 Å. Radii were set at increasing distances to reflect the importance of describing more accurately atoms in the nearer environment, as suggested in the original work by Belongie et al.

As shown in Figure 4, the fingerprint that describes the pocket can be broken down into n fingerprints that describe the environment around each atom of the pocket. Each of these n fingerprints can be broken down into 14 subfingerprints, each of which contains the occurrences of the different atom types in the sphere of radius r_0 (sphere 0) and in each of the 13 spherical layers (sphere 1−13).

1.2. Screening of Protein Pockets. Describing the Query Pocket. Before starting the actual screening of pockets, it is necessary to have a protein in complex with the ligand, i.e., the one for which the inhibition map is to be predicted. The coordinates of this ligand are used to define the query pocket. The original pocket generated by FlapSite (Figure 5a) is shredded by removing first all those points at more than 3 Å from any atom of the ligand (Figure 5b); then the protein atoms at less than 3 Å from this shredded pocket are selected. This approach is useful to select only protein atoms that are at the interface of the pocket occupied by the ligand.

Calculating Binding Pocket Similarity. Once the atoms in the protein are selected to define the query pocket, the fingerprint is generated as described in section 1.1 of the Methods. For each combination of pairs of atoms a_x and a_y of the query pocket and of the database pocket, respectively, a similarity score s is calculated as follows:

$$s_{x,y} = \sum_{i=0}^{13} w_i \times \sum_{p=0}^{17} \frac{(x_{i,p} - y_{i,p})^2}{(x_{i,p} + y_{i,p})} \tag{2}$$

where $x_{i,p}$ and $y_{i,p}$ are the ith and pth elements of the fingerprints that describe the atoms a_x and a_y, i is the sphere number, p is the atom type, and w_i is a weight that depends on the sphere number: $w_i = i^{-2}$ for $i \geq 1$ and $w_i = 1$ for $i = 0$. Different weighting schemes were considered (i^{-1}, i^{-2}, i^{-3}, etc.), and $w_i = i^{-2}$ yielded the best results by using the Ambit data set as a validation set. The selected weighting scheme increases the importance of a good match for atoms showing higher similarity in the nearest environment. The score $s_{x,y}$ is equal to zero if the compared fingerprints are identical.

All the scores s calculated between all possible pairs of atoms, one from each pocket, are stored in a matrix **K**, $n \times m$, where n is the number of fingerprints (atoms) of a protein pocket of the database and m is the number of fingerprints (atoms) of the query pocket. The next step is to find atom equivalences between the atoms in the query pocket and in the database pocket. This is an example of the linear assignment problem (LAP) which, as in the work of Belongie et al.,[20] was solved by implementing the method proposed by Jonker and Volgenant.[25] LAP is solved by calculating permutations over the **K** matrix to minimize the overall score S calculated by summing all the $s_{x,y}$ scores between all pairs of a_x and a_y atoms:

$$S = \sum_x s_{x,y} \tag{3}$$

Therefore, the solution of LAP provides the list of one-to-one atom equivalences between the two pockets and the raw similarity score (S) between two pockets.

Because the number of atoms of a pocket of the database (n) is generally different from the number of atoms of the query pocket (m), the matrix **K** is in most cases a rectangular matrix. Therefore, dummy atoms are added to the smaller pocket to obtain a square matrix, since solving LAP requires a squared matrix. The score for the match of each dummy atom is set to zero (optimal match) if the dummy atom belongs to the query pocket ($n > m$) and to a value $h = 1$ if the dummy atom belongs to a database pocket ($n < m$). Using a dummy atom with a score $s = 0$ is useful for obtaining a good overall match even if the query pocket, because smaller, can only match a portion of the database pocket. For example, when comparing the query pocket of Figure 5b with the pocket of Figure 5a (stored in the database), one would obtain an optimal match. For the alternative case, a database pocket that only partially matches the query pocket is scored unfavorably because this would make the ligand bound to the query pocket unfavorable for binding to the database pocket. Although we found that setting $h = 1$ was useful for predicting inhibition maps for compounds in the

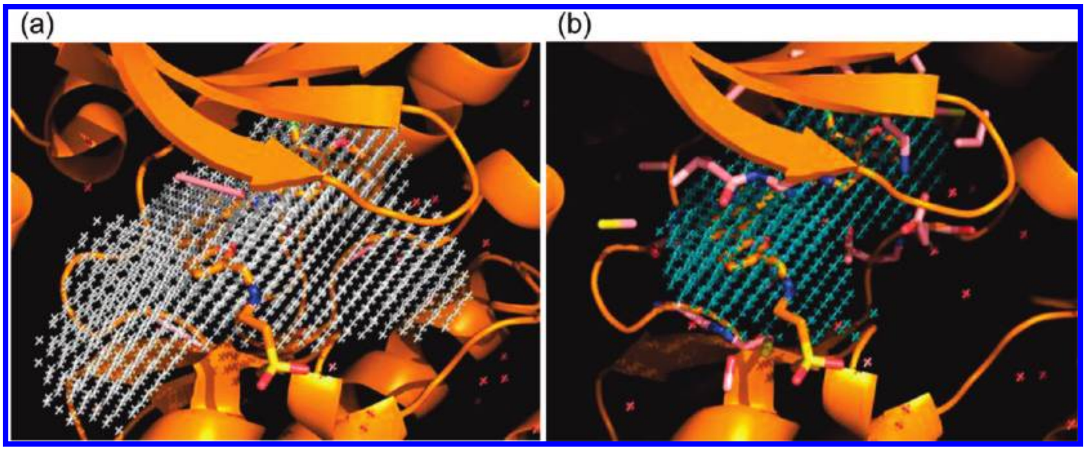


Figure 5. (a) Overall pocket generated by FlapSite for Egfr (1XKK) and (b) the same pocket shredded around the inhibitor Lapatinib. In b, the residues considered to define the pocket are shown in sticks.

Ambit panel, h can also be set to 0 if the interest is more broadly in finding targets with similar subpockets.

1.3. Superposition of the Pockets and Calculation of the Final Similarity Score. The alignment of two pockets was implemented by using the Procrustes algorithm.[23] This method requires a list of atom equivalences, which are obtained from the result of the LAP calculation. However, as suggested by Richmond et al.,[23] who used this method for aligning molecules in 3D, it is critical that this list contains true matches to obtain good alignments. Thus, only atom equivalences with a similarity score s below a threshold are selected. The threshold used is not fixed: we set a minimum threshold with a score of 0.2, and we increment this threshold at intervals of 0.05 units, for a maximum of 10 intervals. The computational cost for the alignment increases with the number of atoms in the list, without necessarily improving the goodness of the alignment. Therefore, we limit the number of atoms for the alignment to 1/10 of the atoms in the database pocket.

To remove geometrically inappropriate atom equivalences, a second filter is applied. All atomic pairs that do not respect atomic distances, as described by Richmond et al.,[22] are discarded. To do so, we calculate for each atom of the identified equivalent pairs the differences in terms of their distances to all of the other atoms in the same pocket: if the difference in the average distance calculated for each atom exceeds 5 Å, that specific atom equivalence is filtered out.

The remaining atom equivalences are used for the alignment, provided that there are at least nine. If there are less than nine atom equivalences, the alignment is not performed because the two pockets are expected to be significantly dissimilar. Given the set of filtered z atom equivalences $\{((a_x)_1 \sim (a_y)_1), ..., ((a_x)_z \sim (a_y)_z)\}$, where a_x and a_y are atoms of the pockets x and y being compared, the rotation matrix to obtain the alignment is calculated using the algorithm described by Rohlf and Slice.[23]

Calculation of the Rotation Matrix. Let M_x and M_y be the matrices containing the coordinates of all the z atoms a_x (query pocket) and a_y (database pocket) of the atom equivalences. We apply a translation T_x to M_x and T_y to M_y so that the coordinates of the two pockets are centered near the origin. The centered coordinates are M'_x and M'_y for M_x and M_y, respectively. We calculate the singular-value decomposition of $M''^{t}_{x}M'_y$, where t denotes the transpose operator. The singular-value decomposition is an object of the form $U\Sigma V^t$, where U and V are 3 × 3 orthogonal matrices, and Σ is a 3 × 3 diagonal matrix of positive eigenvalues. We calculate $H = VU^t$. Last, we calculate the determinant d of $M'^{t}_{x}M_y$, and if $d < 0$, we multiply the last column of V by -1 and then recalculate $H = VU^t$ and terminate. The singular-value decomposition of $M'^{t}_{i}M'_j$ as well as other matrix transformations (transpose of a matrix, matrix multiplication, etc.) were obtained by using the Jama package.[26] To superimpose the pocket y onto pocket x, M_y is first translated by T_y, then rotated by H, and then translated by T_x^{-1}.

Correcting the Similarity Score. The raw similarity score (S) obtained by eq 3 (section 1.2, Methods) can be used to rank pockets; however, we found better results by also taking into account the superposition between the compared pockets. Thus, the RMSD calculated from the superimposed atoms of the pockets is used to obtain the corrected similarity score S':

$$S' = \mathrm{RMSD} \times S \tag{4}$$

If a pocket does not have enough atoms sufficiently similar to the query pocket to obtain an alignment, RMSD cannot be calculated, and a penalty is added:

$$S' = 100 \times S \tag{5}$$

This is particularly useful to penalize possible false positives and to prioritize all pockets that give the best overlay.

It is also important to remember that the rotation matrix obtained as described above is not generated by taking into account all atoms of the two pockets, but only the atoms that are truly equivalent. Eventually, only the RMSD between the coordinates of the atoms used for the alignment is calculated. As a result, pockets that have differences in some regions still have an optimal RMSD if the subset of atoms used for the alignment matches very closely.

Last, because the score S is the sum of the scores s for the matches between all of the atoms composing the two pockets, a further correction is added to normalize pockets of different sizes:

$$S'' = \frac{S'}{n} \tag{6}$$

with n being the number of atoms of the query pocket. For convenience, the similarity score is normalized to a percentage similarity that ranges from 0 to 100, where 100 corresponds to the optimal match.

2. Metrics for the Validation. *The Kinase Panel.* To validate the method presented in this paper, we used K_d data from the Ambit panel, which comprises 38 inhibitors tested on 317 kinases. In order to generate a database of kinase protein pockets, all the PDB structures available for the 317 kinases in the Ambit panel were searched by using SwissProt:[27] this resulted in 1647 PDB structures corresponding to 189 different kinases. Gene names in the Ambit panel were mapped to their corresponding PDB code by using a table extracted from SwissProt UniProtKB using "family:KINASE" and by filtering only results with 3D structure available. The downloadable XML file contains all data for each kinase, including the PDB code.

FlapSite was used to identify pockets for each kinase, and 3957 different pockets were found. Since the procedure was automatic, it is important to note that some of these pockets could not correspond to kinase domains of the target or could correspond to different proteins in complex with the kinase. Furthermore, the Ambit panel reports activity data of the mutant forms for some kinase: data for these mutants were excluded. Out of the 38 inhibitors in the Ambit panel, only 17 are available at least in one crystal structure in PDB. Therefore, for convenience, we limited our study to predict inhibition maps for these 17 inhibitors.

Receiver operating characteristic (ROC) curves were used to assess quantitatively the results of the validation.[28] To obtain ROC curves, true positive rates (tp) are plotted versus false positive rates (fp):

$$\mathrm{tp} = \frac{\text{true targets selected}}{\text{total true targets}} \quad \mathrm{fp} = \frac{\text{decoys selected}}{\text{total decoys}} \tag{7}$$

A true target is here defined as a kinase that is inhibited with $K_d < 10\ \mu$M, and a decoy is a kinase that is inhibited with $K_d > 10\ \mu$M. It is important to highlight that many decoys are kinases still inhibited in the low micromolar range, which makes the benchmark particularly challenging.

The ideal ROC curve yields an area under the curve (AUC) = 1 because all of the true targets rank at the top of the list, whereas a random ROC curve yields AUC = 0.5. However, because the interest is in understanding whether the kinases that are at the top of the ranking are inhibited by a given compound, the metrics used for the validation are the ROC enrichment 10% (ROC $\mathrm{Enr}_{10\%}$), which is the true positive rate at 10% of the false positive rate. In other words, ROC $\mathrm{Enr}_{10\%}$ is the percentage of true targets predicted when 10% of the decoys are retrieved. The ideal ROC $\mathrm{Enr}_{10\%}$ value is 1, whereas a random ROC curve yields a ROC $\mathrm{Enr}_{10\%}$ = 0.1. The entire ROC curves provide a useful visual guide of the performance of the method. For quantitative purpose, ROC $\mathrm{Enr}_{10\%}$ is more suited to assess early enrichment; AUC provides a measure of the performance of the method across the entire ranked database.

The Protein Universe. In addition to the Ambit kinase database of pockets, a database of pockets derived from all the protein structures available from the PDB (November 2009) was also generated. After removal of the structures with pockets that were either too large (>4000 Å^2) or too small (<300 Å^2), 199 652 pockets corresponding to 46 186 different PDB structures were produced. Both the Ambit kinase and the whole PDB databases of pockets were screened starting from the 17 kinase inhibitors of the Ambit panel. The results of these screenings will be discussed on a case by case basis by using experimental data collected from literature and by analyzing the 3D alignment to support the validity of the predictions. The screening of the whole PDB database of nearly 200 K pockets requires about 50 h/CPU, which makes the method presented in this paper sufficiently fast, with the aid of a computer cluster, for routine use.

RESULTS

1. Predicting Inhibition Maps for Kinase Inhibitors by Pocket Similarity. *ROC Enrichment.* To assess quantitatively the accuracy of the method presented in this paper, ROC curves were generated for the 17 inhibitors (Supporting Information Figure 1). AUC and ROC $\mathrm{Enr}_{10\%}$ values are reported in Table 2. When more than one PDB structure was available with the same inhibitor bound, we selected the target with the highest binding affinity and better resolution (reported in column PDB of Table 2). Table 2 shows that, on average, 37% of the kinases are retrieved at ROC $\mathrm{Enr}_{10\%}$. The best result was obtained with Lapatinib ($\mathrm{Enr}_{10\%} = 0.75$): ErbB4 and Egfr ranked first and second, respectively, and Stk10 ranked fifth. The fourth most efficacious target of Lapatinib, ErbB2, ranked only 57th. This is not a deficiency of the method but is to be ascribed to the fact that the only two PDB structures available for ErbB2 (1S78 and 1NZ8) do not include the catalytic kinase domain, but only the extracellular domain. Therefore, Lapatinib would have yielded $\mathrm{Enr}_{10\%} = 1$ if these ErbB2 structures had not been included in the benchmark. Because the selection of kinases extracted from the PDB was based on the general kinase annotation reported on SwissProt, this mismatch in the generation of the database was not preventable, and we cannot rule out that other structures have a similar problem.

To assess how the choice of a particular PDB structure as a query affects the results of the enrichment, calculations were also performed by using all of the other available PDB structures for each of the inhibitors under investigation. For Imatinib, eight different PDB complexes were used, yielding ROC $\mathrm{Enr}_{10\%}$ values ranging from 0.23 to 0.38. The three complexes that gave the lowest enrichment (all ROC $\mathrm{Enr}_{10\%}$ = 0.23) were 3HEC (p38), 2OIQ (Src), and 2PL0 (Lck), whereas those that produced the highest enrichment (all ROC $\mathrm{Enr}_{10\%}$ = 0.38) were 3GVU (Abl2) and 1OPJ (Abl1). The use of a kinase that is not a *true target* for the inhibitor under study (p38 and Src, $K_d > 10\ \mu$M) causes poorer results. This can be partly ascribed to the fact that the pocket residues around the ligand are not optimal for binding. However, Imatinib shows a K_d = 40 nM for Lck. In general, ROC curves show that the choice of the PDB structure does not affect results in a significant way. The results are different if available nonkinase structures are used as query pockets. For example, crystal structures for Roscovitine and Flavopiridol in complex with nonkinase targets are also available in the PDB: 1YGK, which is a Pyridoxal kinase (PDXK) bound to Roscovitine; 1C8K and 1E1Y, which are glycogen phosphorylase b (GPb) bound to flavopiridol and glycogen phosphorylase a (GPa) bound to flavopiridol and glucose,

Table 2. ROC AUC and ROC $\mathrm{Enr}_{10\%}$ Calculated Using the Method Presented in This Paper (Pocket Similarity) on a Training Set of 17 Inhibitors[a]

					pocket similarity		MDL public keys		flap	
inhibitor	kinase target	PDB	N actives	K_d (nM)	ROC AUC	ROC $\mathrm{Enr}_{10\%}$	ROC $\mathrm{Enr}_{10\%}$ (a)	ROC $\mathrm{Enr}_{10\%}$ (b)	ROC $\mathrm{Enr}_{10\%}$ (a)	ROC $\mathrm{Enr}_{10\%}$ (b)
AMG-706	Vegfr2	3EFL	17	26	0.69	0.41	1.00	0.50	0.60	0.43
BIRB-796	Ptk2b	3FZS	37	990	0.63	0.41	0.66	0.06	0.45	0.23
CP-690550	Jak2	3FUP	4	5	0.71	0.25	0.14	0.43	0.43	0.00
Dasatinib	Abl1	2GQG	47	0.53	0.7	0.47	0.38	0.07	0.32	0.14
Erlotinib	Egfr	1M17	21	0.67	0.67	0.38	0.60	0.15	0.64	0.27
Flavopiridol	Cdk9	3BLR	32	6.4	0.54	0.19	0.07	0.15	0.16	0.16
Gefitinib	Egfr(G719S)	2ITO	15	1.1	0.69	0.33	0.27	0.52	0.76	0.21
Imatinib	Abl2	3GVU	13	10	0.64	0.38	0.59	0.63	0.41	0.70
Lapatinib	ErbB4	3BBT	4	54	0.9	0.75	0.40	0.60	1.00	0.20
LY-333531	Pdpk1	1UU3	26	700	0.5	0.23	0.09	0.52	0.26	0.30
Roscovitine	Cdk5	1UNL	7	1900	0.72	0.57	0.10	0.00	0.37	0.36
SB-203580	p38-alpha	2EWA	19	12	0.62	0.47	0.32	0.27	0.61	0.15
Sorafenib	B-Raf	1UWH	38	230	0.6	0.34	0.72	0.07	0.63	0.25
Staurosporine	Chek1	1NVR	124	3.2	0.6	0.30	0.27	0.92	0.48	0.69
Sunitinib	Kit	3G0F	81	0.37	0.57	0.27	0.26	0.28	0.18	0.97
VX-680	Abl1(H396P)	2F4J	66	9.1	0.55	0.24	0.53	0.30	0.26	0.45
ZD-6474	Ret	2IVU	45	34	0.61	0.31	0.62	0.11	0.43	0.27
average					0.64	0.37	0.41	0.33	0.47	0.34

[a] ROC $\mathrm{Enr}_{10\%}$'s are reported from MDL public keys and Flap fingerprints on a training set of (a) 37 or (b) 16 inhibitors. The training set includes inhibitors in the Ambit panel except the one being predicted.

respectively. 1YGK yielded only Enr10% = 0.14 as opposed to Enr10% = 0.57 obtained by using 1UNL (Roscovitine-Cdk5). Roscovitine has micromolar affinity for PDXK and binds to PDXK in a conformation different from that adopted in protein kinases. 1E1Y and 1C8K did not yield any significant enrichment (ROC $\mathrm{Enr}_{10\%}$ = 0), as opposed to 3BLR (ROC $\mathrm{Enr}_{10\%}$ 0.19), which is a flavopiridol−Cdk9/cyclin T1 complex. In these crystal structures, Flavopiridol has a similar binding conformation. However, while Flavopiridol shows a K_d of 6.4 nM against cdk9, it inhibits GPb with an IC_{50} of 15.5 μM.[29] The inhibition is synergistic with glucose, resulting in a reduction of IC_{50} for Flavopiridol to 2.3 μM.

Inhibition maps. Figure 6 shows the predicted inhibition maps for the 17 Ambit inhibitors. The map was built by including all targets with a percentage similarity > 75%. Targets with a K_d < 10 μM (true targets) are linked to the central node (the target used as a query) by a black line, whereas targets with K_d > 10 μM (decoys) are not linked to the central node. Because each inhibitor was processed on the entire PDB, other targets (kinase ot not) for which K_d values are not available are also reported in Figure 6. These are linked to the central node by an orange line. The size of the node reflects the degree of similarity to the query pocket, with a larger node indicating a higher degree of similarity. Each node is colored by kinase family.

By analyzing the decoys reported in Figure 6, targets with K_d in the low micromolar range, but still >10 μM, were found. For example, among the false positive targets predicted for Imatinib there are Src, Vgfr2, B-Raf V600E mutant, and Mk14. The case of Src has been discussed in the previous section, and the retrieved structure (3G6G) shows Src in its inactive DFG-out conformation bound to Imatinib. The affinity of Imatinib for Vgfr2 is 10.7 μM[30] and is 3.3 μM for the B-Raf V600E mutant, while it is >10 μM for B-Raf1. Last, Mk14, which corresponds to the 1KV2 PDB structure of p38-α kinase with Imatinib (K_d > 10 μM), suggests that this is a case similar to that of Src. These examples show that the actual performance of the method is underestimated by considering the ROC $\mathrm{Enr}_{10\%}$ value alone.

For some inhibitors, we predicted as true targets, kinases that are not included in the Ambit panel. For example, Irak4, Fak, and Snf were predicted as possible targets of Sunitinib, with Irak4 and Fak having a very high similarity to the query pocket. By investigating additional literature data,[31] we found that Sunitinib does inhibit Irak4 (28% inhibition at 0.3 μM and 89% inhibition at 10 μM) and Fak (32% inhibition at 0.3 μM and 94% inhibition at 10 μM), whereas no information could be retrieved regarding Snf kinase.

Several high ranking targets predicted for Flavopiridol are not included in the Ambit panel, such as casein kinase 2a (Csk2a), Pask, dual specificity tyrosine−phosphorilation regulated kinase 1A (Dyr1a), Fus3 (yeast MAP kinase), Sky1 (yeast SR kinase), Cdc2h (*P. Falciparum* Cdk2), and Cgh2 (Cdk6). The link between Csk2a and Flavopiridol is reinforced by the similarity between Flavopiridol and Csk2a inhibitors, which have a flavonoid structure, as the one in the 1M2P PDB structure. This is an interesting example for showing that findings from this study may be used to initiate the design of new inhibitors from the structure of inhibitors that hit related targets. The 84 nM affinity of Flavopiridol for Dyr1b reported in the Ambit panel may explain the high ranking of Dyr1a.

Selectivity Score. The inhibition maps in Figure 6 highlight that some compounds are predicted to be much more promiscuous than others. Since different inhibitors bind to the same target and still have very different promiscuity profiles, predicting the promiscuity of a compound involves understanding the similarity among targets only in those regions of the pocket that interact directly with the inhibitor. The original paper containing the K_d values against the Ambit panel reports an experimental selectivity score for each of the 38 compounds in the panel by dividing the number of kinases that bind with a K_d < 3 μM by the total number of kinases tested (excluding mutants).

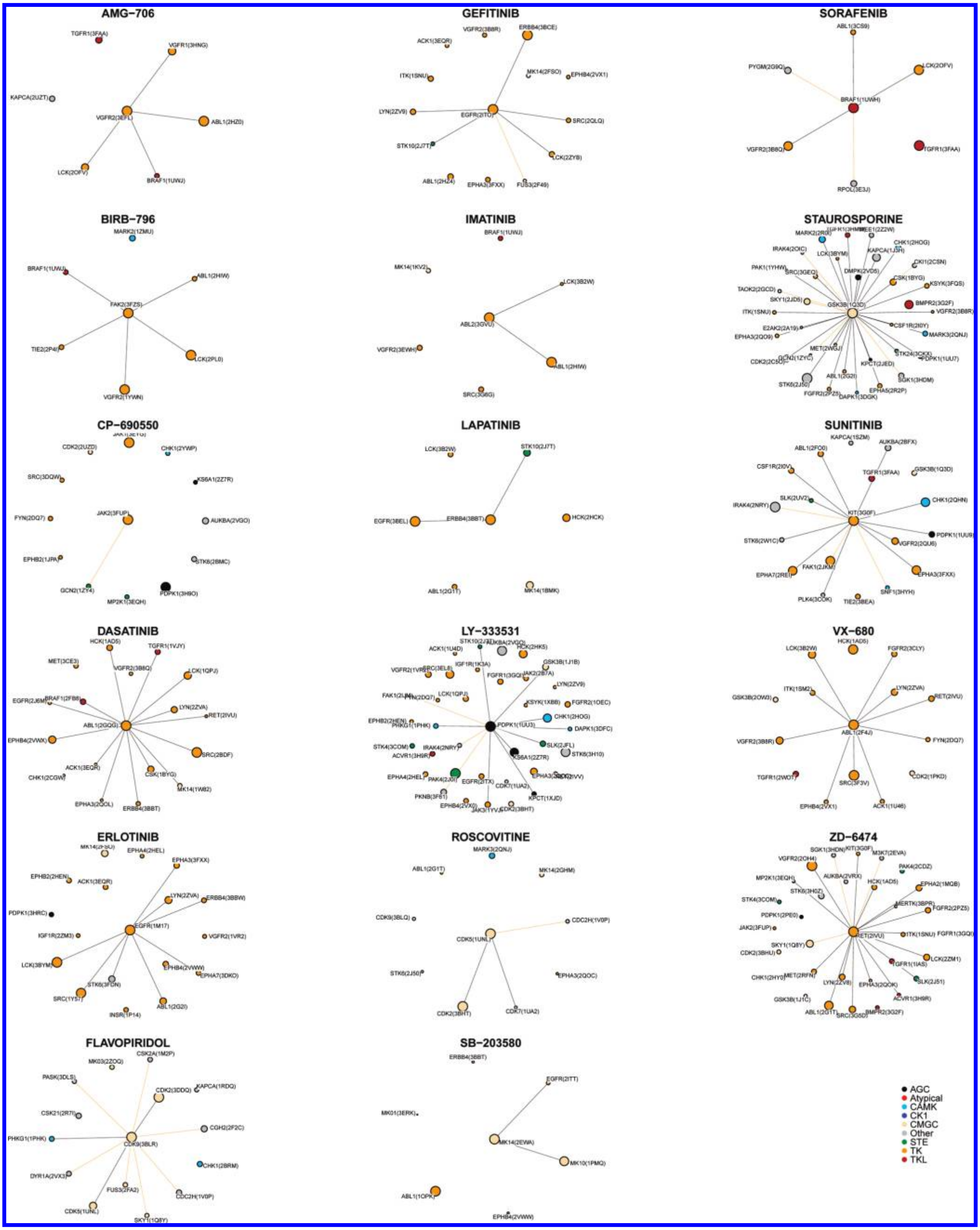


Figure 6. Inhibition maps for 17 kinase inhibitors in the Ambit panel. All targets with a percentage similarity > 75% are reported. Targets with a K_d < 10 μM (true targets) are linked to the central node (the target used as a query) by a black line, whereas targets with K_d > 10 μM (decoys) are not linked to the central node. Nodes linked to the central node by an orange line correspond to (kinase or nonkinase) targets for which affinity data are not available. The size of the node reflects the degree of similarity to the query pocket, with a larger node indicating a higher degree of similarity. Each node is colored by kinase family (see legend).

To understand whether our computational method can be used to rank different compounds by their selectivity, selectivity scores predicted by our method were calculated by counting the number of targets with a percentage

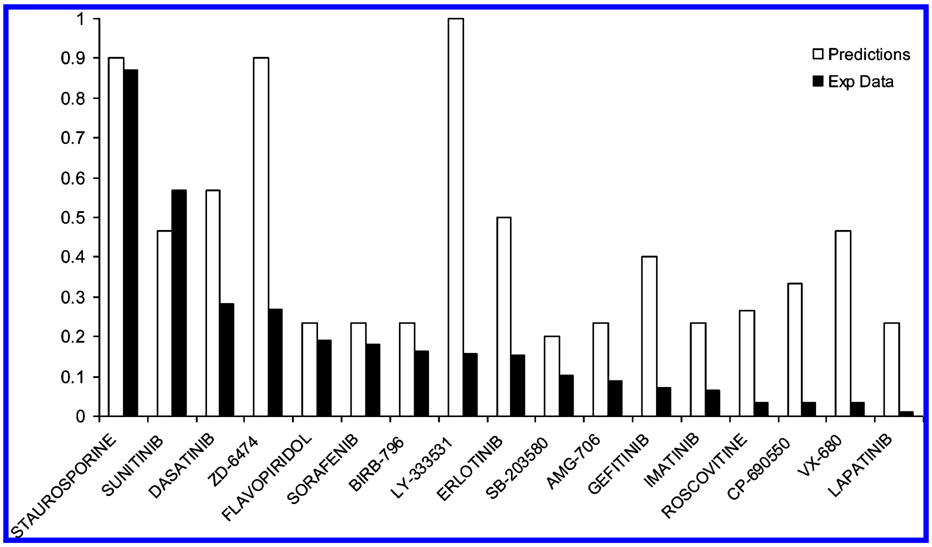


Figure 7. Promiscuity profile of the 17 inhibitors in the Ambit panel using predicted data and experimental data.

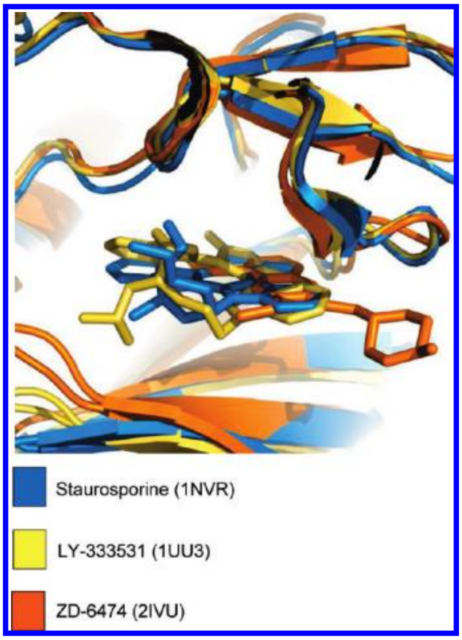


Figure 8. Compounds predicted to be more promiscuous. The promiscuity score predicted for Staurosporine, LY-333531, and ZD-6474 is high because in all cases the same highly conserved ATP binding subregions are occupied. However, the actual promiscuity is influenced also by other factors, such as ligand flexibility, as in the case of LY-333531.

similarity > 75% and dividing this by the number of kinases in the Ambit panel for which a 3D structure is available (i.e., 189 structures). The experimental and predicted selectivity scores are plotted in Figure 7. Both have been normalized so that the compound predicted to be the most promiscuous has a score of 1.

The compounds predicted to be more promiscuous (Staurosporine, LY-333531, and ZD-6474) are those that occupy the more conserved ATP binding regions across all kinases (namely, the adenine, the ribose, and the phosphate binding regions),[32] as shown in Figure 8. However, while this prediction is correct for Staurosporine, the Ambit panel shows that LY-333531 (Ruboxistaurin) is much more selective, and this has been attributed to its less planar structure.[33] Experimental data also show that ZD-6474 is not as promiscuous as predicted; however, it still ranks among more promiscuous compounds. The actual promiscuity can also be influenced by other factors, such as conformational changes of the protein and of the ligand flexibility, and such effects are not taken into account by this method.

A limited number of studies conducted on internal data, on compounds of the same series, provides a better selectivity score performance. This is likely because ligand similarity is high; therefore differences in selectivity are influenced almost exclusively by the pocket differences in the same portion of the binding pockets occupied by the ligand. This improvement needs to be validated on a larger number of cases.

2. Predicting Inhibition Maps for Kinase Inhibitors by Ligand Similarity. Ligand-based methods are also very useful for discovering inhibitors with the desired selectivity profile, and important results have been reported by using these approaches.[14,34] As opposed to target-based methods, ligand-based methods do not require the 3D structure of the protein-inhibitor complex; however experimental activity data are needed as a training set.

To understand strengths and weaknesses of our target-based method compared to ligand-based methods, a pairwise ligand similarity matrix was calculated across all the Ambit 17 ligands. Starting from the hypothesis that similar compounds are likely to hit the same targets, a network of targets is generated using the similarity of their ligands. For each of the 17 inhibitors, a list of putative targets was generated by considering the targets inhibited (with $K_d < 10\ \mu$M) by compounds sorted by decreasing similarity. The list of targets was ordered starting from the targets of the most similar compound and ranking the targets with higher affinity first. The procedure was repeated by listing the targets of the second most similar compound, and without repeating targets already listed. The obtained ranked list of targets was used to calculate ROC curves in analogy with those obtained by binding pocket similarity.

MDL public keys were selected as 2D fingerprints within Pipeline Pilot[35] by using the “Molecular Fingerprints” and “Fingerprint Similarity N×N” components to obtain the similarity matrix for the 17 inhibitors. FLAP was used to generate four-point 3D pharmacophores for each ligand. The FLAP procedure analyzes the space around the ligand by using GRID molecular interaction fields (MIFs) obtained by running a limited series of chemical probes (H, DRY, N1, O). The probes enable the identification of energetically favorable and unfavorable interactions. The information contained in the GRID-MIFs is then condensed into fewer pharmacophoric points. All possible energetically favorable arrangements of four pharmacophoric points are then produced and encoded in a fingerprint. The conformational flexibility and the shape of the ligand are also taken into account. Various conformers are generated by the program (RMSD of 0.2 Å between two conformers up to maximum of 25 conformers), and MIFs are calculated for each conformer. The similarity score between different conformers of the compared compounds is calculated by identifying the best superposition of quadruplets, directly comparing the volumes of the oriented MIFs for each probes being used individually, and then combining them in order to produce probe-combination scores. The probe-combination score, defined as the Distance Score, was used to rank compounds.

Table 2 reports the ROC $\mathrm{Enr}_{10\%}$ obtained by these 2D and 3D ligand-based methods. On average, ROC $\mathrm{Enr}_{10\%} = 0.34$ was obtained using 3D fingerprints and ROC $\mathrm{Enr}_{10\%} = 0.33$ using 2D fingerprints. Because this type of approach depends on the used training set, the calculations using the whole set of 38 inhibitors were repeated. This yielded an average ROC $\mathrm{Enr}_{10\%} = 0.47$ using 3D fingerprints and ROC $\mathrm{Enr}_{10\%} = 0.41$ using 2D fingerprints. The use of a larger training set allows a better prediction. For example, using FLAP we obtained ROC $\mathrm{Enr}_{10\%} = 1$ for Lapatinib because the most similar ligand is CP-724714, which hits the same targets of Lapatinib (Egfr, ErbB2, ErbB4) in addition to Mlck and Cdk11 and Cdk8. The smaller training set of 17 inhibitors does not include a ligand so similar, and thus the obtained ROC $\mathrm{Enr}_{10\%}$ value was much lower (0.20). However, this effect was not consistent for all inhibitors, and we observed either enhancements or worsening in the predictions depending on the specific compound.

Despite the small ligand training set, these results show that the performance of ligand-based methods is comparable to that of target-based method. Therefore target- and ligand-based approaches have great potential, and both could provide useful information. The combination of these two different approaches might provide an improved overall success rate in polypharmacology prediction. However while ligand-based methods are particularly suitable for compound screening, target-based methods are better suited for the optimization of the compound selectivity profile.

3. Case Studies: Exploring the Protein Universe. In the previous section, we have shown the application of the method to the prediction and rationalization of the inhibition map of kinase inhibitors: a high degree of accuracy is obtained when suitable PDB structures are available. Starting from kinases as queries, we identified possible off-targets with low sequence identity to the query, as shown in Figure 6. However, all targets had the kinase fold. For that reason, two other data sets were considered for the investigation of the pocket similarity in nonhomologous proteins (i.e., showing no sequence/fold similarity): (1) ATP binding proteins and (2) proteins bound to 10 compounds (including various marketed drugs) of biological relevance, hitting targets outside the kinase superfamily. While a percentage similarity score of 75 was used in the case of kinases, a lower threshold of 50 was selected for this section. This choice was motivated by the fact that here the focus is on finding target similarities at the 3D level, rather than on predicting all targets of a given inhibitor.

3.1. ATP. The goal of this first case study was to investigate the possibility of linking targets with different folds, but binding the same ligand, by looking at the similarity of their binding pockets. The use of ATP as a ligand is particularly suitable because numerous PDB structures are available for proteins that bind either ATP or related nucleotide ligands. Six representative nonhomologous proteins bound to ATP were selected: Pdk1 kinase (3HRC), phosphoenolpyruvate carboxykinase (2OLR), actin (2GWK), aspartate carbamoyltransferase (2YWW), tryptophanyl-tRNA synthetase (1M83), and pyridoxal kinase (3FHY). Starting from each of these six complexes listed above, the whole PDB pocket database was screened to identify similar targets.

As shown in Figure 9, the usage of pyridoxal kinase (PDXK) and aspartate carbamoyltransferase (PYRY) as query pockets led to the retrieval of only alternative PDB structures of the same target (for that reason not reported in the inhibition maps of Figure 9). In the case of actin as a query pocket, few variants were found, but although some of these have low sequence similarity, all of the retrieved targets fall into the actin protein family. In the case of tryptophanyl-tRNA synthetase, only one similar pocket was retrieved—pantothenate synthase—which has only 6% sequence identity with the query and a completely different fold, but it is known to bind ATP.

The most interesting case is that of *E. coli* phosphoenolpyruvate carboxykinase (PCKA), which is a lyase that catalyzes the carboxylation of phosphoenolpyruvate to oxaloacetate. Starting from the PDB structure (2OLR) of PCKA bound to ATP, several targets with nearly identical binding pockets, but completely different folds, were retrieved. The highest ranking target is the GTP-binding nuclear protein RAN (1WA5); the second is a cystic fibrosis transmembrane conductance regulator (1XMJ, membrane protein, hydrolase); the third is the human small GTPase Rab7; the fourth is a variant of PCKA (PCKGC, 3DT4) which binds GTP rather than ATP as in the query; the fifth is the catalytic domain of a Ras protein bound to GDP (1Q21, oncogene protein), and the sixth is the ATP-dependent RNA helicase MSS116 (3I5Y).

The most interesting observation to make is about the alignment obtained by the program presented in this paper. The obtained alignment highlights not only an optimal overlap between atoms of the proteins at the pocket level but also a good overlay of the corresponding ligands (Figure 10), which were not taken into account to produce the alignment. The ligands bound to these high ranking targets are ATP, GTP, GDP, and the ATP analogue AMP-PNP. In all cases, the coordinates of the phosphate atoms had the highest degree of overlay compared to the rest of the structure. In one case (1XMJ, Figure 10), the alignment shows that the adenine portion of the ligands points to different directions, whereas the phosphate is overlaid. This finding shows that when two pockets are very similar in a subpocket, a high similarity can still be revealed even if the ligand changes its conformation. Generally, this high similarity is driven by the very low RMSD, which is used to correct the similarity score.

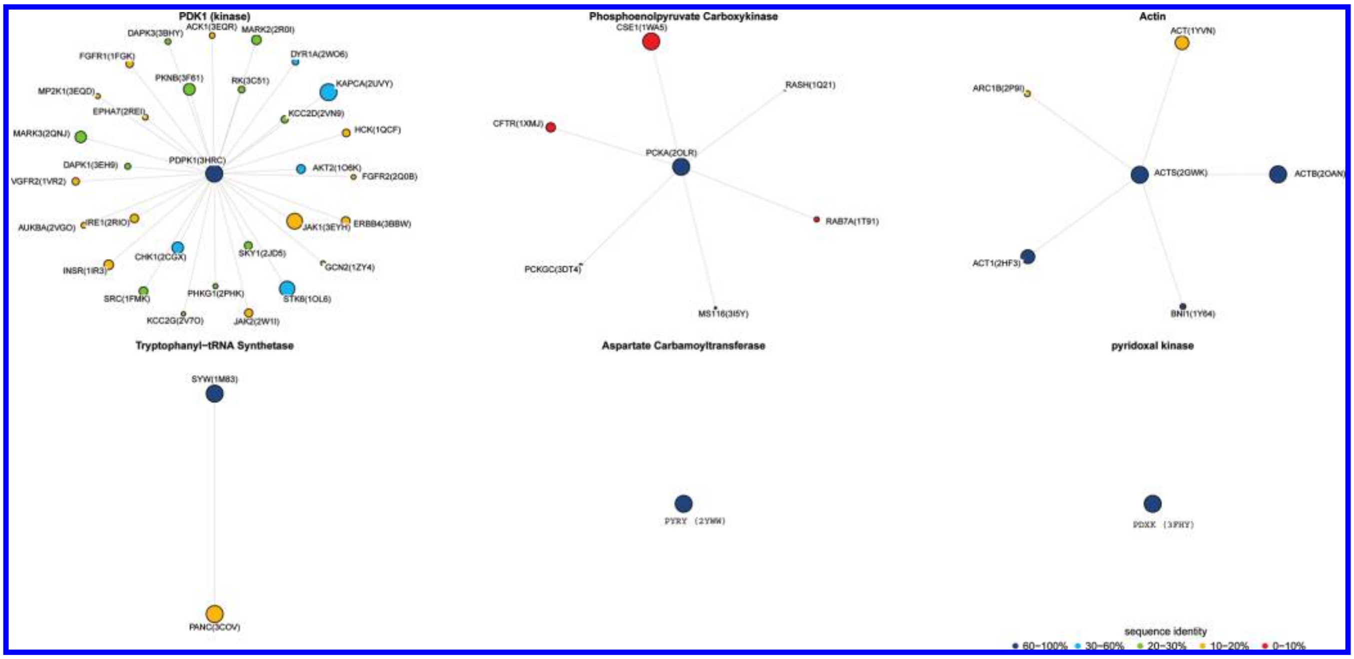


Figure 9. Inhibition maps of ATP binding proteins. Six representative nonhomologous proteins bound to ATP were selected: Pdk1 kinase (3HRC), phosphoenolpyruvate carboxykinase (2OLR), actin (2GWK), aspartate carbamoyltransferase (2YWW), tryptophanyl-tRNA synthetase (1M83), and pyridoxal kinase (3FHY). Starting from each of these six complexes listed above, the whole PDB pocket database was screened to identify similar targets. The size of the node reflects the degree of similarity to the query pocket, with a larger node indicating a higher degree of similarity. Each node is colored by sequence identity with respect to the query (see legend).

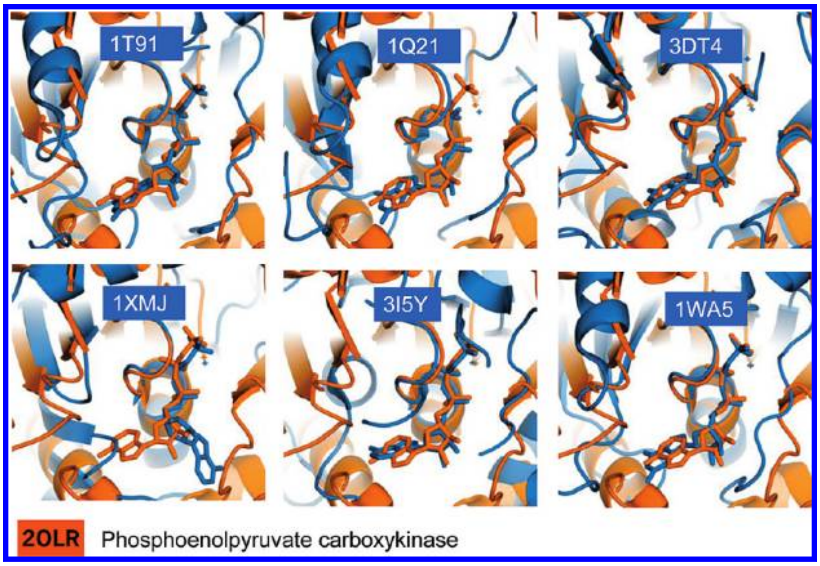


Figure 10. Proteins with completely different folds (in cyan) and nearly identical binding sites overlaid by the program presented in this paper to phosphoenolpyruvate carboxykinase (2OLR, orange). The case of 1XMJ is interesting because this target scored in top positions although the structural similarity is limited to the phosphate binding region.

A more exhaustive study was performed by using ROC curves as a metric for assessing the capability of the pocket similarity method in retrieving an ATP-binding pocket starting from the six queries described above. This study requires a classification of the PDB structures as *ATP-binding proteins* (*true target*) or *ATP nonbinding proteins (decoys)*. The ATP binding proteins were defined as all the unique targets having at least one PDB structure in complex with ATP (234 unique proteins for a total of 469 PDB structures in complex with ATP). All the other PDB crystal structures corresponding to the same set of 234 targets were also included in the set of PDB structures of ATP-binding proteins. This matching of targets was performed by using the SwissProt nomenclature for the PDB proteins. This classification could not be exhaustive as the majority of known ATP binding proteins do not have a PDB structure with ATP bound in the complex. In addition, many proteins bind ligands related to ATP (adenine or nucleotide ligands), thus challenging the current experiment. By taking into account only unique targets, 234 proteins were classified as *ATP-binding proteins*, and 7575 were classified as *ATP nonbinding proteins*. For each target, all the pockets were generated by FlapSite as previously described.

ROC enrichment curves (reported as Supporting Information Figure 2) were calculated starting from each of the six query targets of Figure 9: on average, 30% of *ATP-binding proteins* were retrieved at ROC $\mathrm{Enr}_{10\%}$. This result was consistent regardless of the target used as a query. A

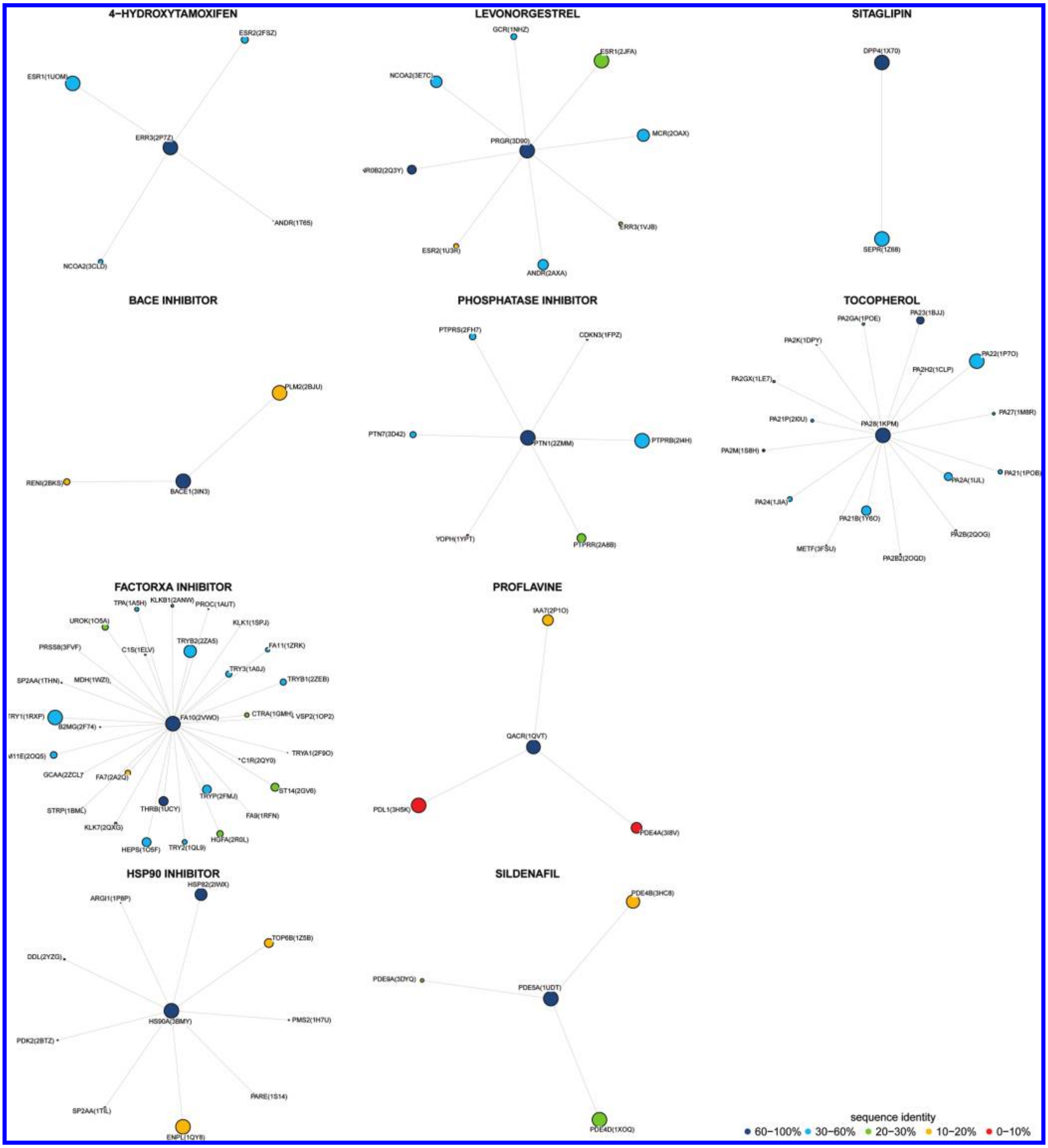


Figure 11. Inhibition maps of 10 biologically active compounds. Starting from each of the 10 complexes, the whole PDB pocket database was screened to identify similar targets. The size of the node reflects the degree of similarity to the query pocket (central node), with a larger node indicating a higher degree of similarity. Each node is colored by sequence identity (see legend) with respect to the query.

comparison of this result with that obtained by sequence similarity (ClustalW[36]) shows that the sequence-based enrichment was poorer (ROC $\mathrm{Enr}_{10\%}$ 15%) in the case of actin, phosphoenolpyruvate, and pyridoxal kinase, whereas it was equally as good as that by pocket similarity (ROC $\mathrm{Enr}_{10\%}$ 27%) in all other cases.

3.2. Marketed Drugs and Other Biologically Active Compounds. In the second case study, a data set of PDB complexes for 10 biologically active compounds (some of which are marketed drugs) was used as a set of queries to predict their inhibition maps considering the entire PDB (Figure 11). The targets associated with the active compounds cover a variety of protein families, such as protease (Factor XA, BACE, DPP4), phosphatase, phospholipase, nuclear receptor (estrogen and progesteron receptor), transcription regulator, phosphodiesterase, and Hsp90. Kinases were excluded, as they have already been deeply studied.

Proteins with similar functions were retrieved in top ranking positions for the following four query targets: (1) From Tocopherol bound to phospholipase A2 and (2) from

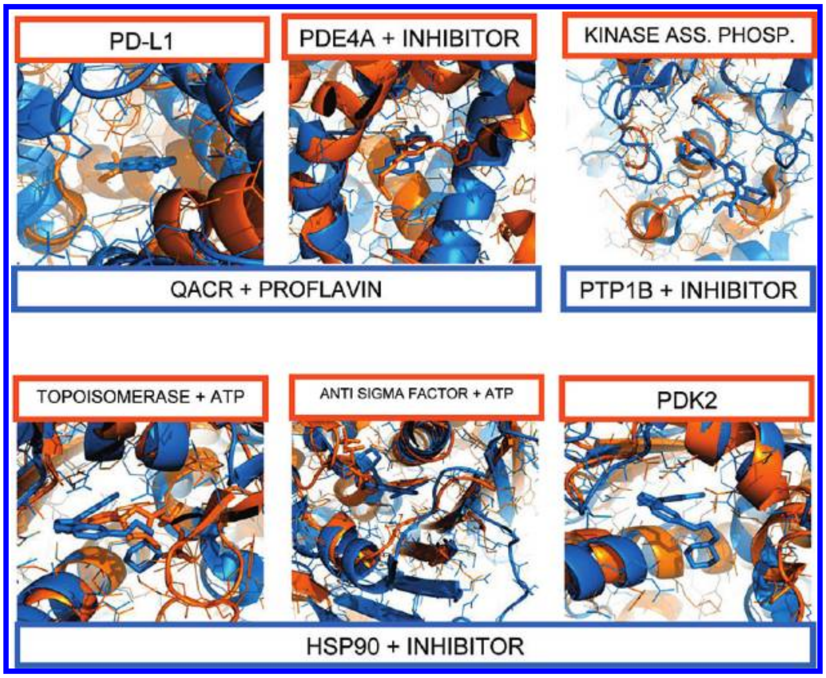


Figure 12. The 3D alignment at the binding site level of targets with different folds.

Sildenafil bound to phosphodiesterase A5 (PDE5A), only variants of the query target were retrieved. Experimentally, Sildenafil has an IC_{50} of 7 μM[37] for PDE4D, the most similar target to PDE5A (Figure 11). (3) Likewise, the query Levonogestrel bound to the progesterone receptor identified other related targets: androgen receptor (ANDR, PDB file 2AXA), mineralocorticoid receptor (MCR, 2OAX), glucocorticoid receptor (GCR, 1NHZ), estrogen receptor (ESR1, 2JFA), which are also all experimentally known targets of Levonorgestrel.[38] (4) The 4-hydroxytamoxifen query, which is an inhibitor of the estrogen receptor, resulted in the identification of other estrogen receptors, in addition to the glucocorticoid and the androgen receptors.

Proteins with different functions were retrieved in top ranking positions with the other query targets: (5) Among the top ranking targets of the PDB structure 3BMY (Hsp90) a topoisomerase (1Z5B), a Pdk2 (protein kinase, 2BTZ), and an antisigma F factor (transcriptase, 1TIL) were found. Currently, we do not know if these proteins bind the inhibitor cocrystallized with Hsp90. However, all these proteins are known to bind ATP, like Hsp90. Therefore, similarity among these pockets is reasonable. Moreover, the structural alignment of these PDB structures with the query pocket is good (Figure 12). (6) Using proflavin bound to the multidrug transcriptional repressor QACR (1QVT), we retrieved PD-L1 (1H5K), PDE4A (3I8V), and a transport inhibitor response protein (2P1O). The similarity with PDE4A is particularly interesting, since the ligand cocrystallized with PDE4A is very well superimposed to proflavin, as shown in Figure 12. (7) An additional case of targets with similar binding pockets and different folds is that of a protein tyrosine phosphatase 1B (2ZMM) and a kinase associated phosphatase (1FPZ), as illustrated in Figure 12. (8) The use of the PDB structure 3IN3 of BACE resulted only in the retrieval of Renin and *P. falciparum* Plasmepsin-2 as possible off-targets, which are well-known possible off-targets of BACE inhibitors. (9) The other proteases used as queries found only related targets (e.g., factor Xa, yielded a variety of serine proteases with the same fold of the query). (10) The use of antidiabetic Sitaglipin bound to dipeptidyl peptidase 4 as a query retrieved Seprase as a target. Sitaglipin is known to be very selective to DPP4. However off-target effects of other DPP4 inhibitors associated to Seprase, DPP8, and DPP9 are well-known. DPP8 and DPP9 are not available in the PDB database.

CONCLUSIONS

In this paper, we have presented a method for predicting polypharmacology starting from the PDB crystal structure of a target in complex with the ligand of interest. We have found that the method is suitable for rationalizing the selectivity profile of kinase inhibitors, by identifying the most likely targets according to their similarity to the query target at the binding site level. The method was validated using a subset of 17 kinase inhibitors profiled against the Ambit kinase panel (189 kinases) and was demonstrated to be effective in retrieving targets showing a $K_d < 10$ μM, also in cases where sequence identity is low.

The importance of examining ligand binding sites and not only protein sequence similarity when assessing similarity between two targets is important in identifying similar targets with different folds but that are unexpectedly similar at the binding site level. The proposed fingerprint description of the binding site can derive useful chemogenomic relationships among proteins not belonging to the same family. By using ATP as a ligand and 10 ligands bound to pharmaceutically relevant targets, we were able to detect ligand cross-reactivity or local similarities at ligand subpockets for proteins with totally different folds and functions. In fact, the screening of the whole PDB pocket database using ATP as a ligand leads to the identification of other known ATP-binding targets with different folds from the one used as a query. The visual inspection of the database targets aligned to the query enables the identification of which is the portion of the pockets that is similar; e.g., local similarities are found only between the phosphate binding regions of ATP/GTP binding proteins, in which the nucleotide portion of the ligand is not overlaid. The detection of high local similarity is considered useful in (de novo) design applications. For example, the retrieval of a portion of a ligand (fragment) accommodated in a particular subpocket that shares similarity with the pocket of interest can inspire the design of novel ligands/fragments to synthesize/screen.

Finally, the proposed structure-based approach, in which protein pockets are compared, combined with ligand-based approaches, in which chemical similarity among ligands is used, could provide an improved overall success rate in polypharmacology prediction.

Supporting Information Available: Figure 1 contains ROC enrichment curves (calculated as described by eq 7) for the 17 inhibitors of Figure 6 by the proposed pocket similarity method (validation study described in section 1 of the Methods). Figure 2 contains ROC enrichment curves (calculated as described by eq 7) for the six nonhomologous proteins bound to ATP in Figure 9 (validation study described in section 3.1 of the Results). This information is available free of charge via Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

(2) Buchdunger, E.; Cioffi, C. L.; Law, N.; Stover, D.; Ohno-Jones, S.; Druker, B. J.; Lydon, N. B. Abl protein-tyrosine kinase inhibitor STI571 inhibits in vitro signal transduction mediated by c-kit and platelet-derived growth factor receptors. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 139–145.

(3) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.

(4) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recog.* **2010**, *23*, 209–219.

(5) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-Throughput Virtual Screening of Proteins Using GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2010**, *50*, 155–169.

(6) Goodford, P. J. A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(7) Schmitt, S.; Kuhn, D.; Klebe, G. A. New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(8) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(9) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33*, W337–W341.

(10) Jambon, M.; Imberty, A.; Delage, G. Geourjon, C.A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 137–145.

(11) Kinnings, S. L.; Jackson, R. M. Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family. *J. Chem. Inf. Model.* **2009**, *49*, 318–329.

(12) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small-molecule protein-binding sites. *Bioinformatics* **2008**, *24*, i105–i111.

(13) Konrat, R. The protein meta-structure: a novel concept for chemical and molecular biology. *Cell. Mol. Life Sci.* **2009**, *66*, 3625–3639.

(14) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. *J. Med. Chem.* **2004**, *47*, 550–557.

(15) Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat. Chem. Biol.* **2008**, *4*, 691–699.

(16) Levinson, N. M.; Kuchment, O.; Shen, K.; Young, M. A.; Koldobskiy, M.; Karplus, M.; Cole, P. A.; Kuriyan, J. A Src-like inactive conformation in the abl tyrosine kinase domain. *PLoS Biol.* **2006**, *4*, 753–767.

(17) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236–4243.

(18) Seeliger, M. A.; Nagar, B.; Frank, F.; Cao, X.; Henderson, M. N.; Kuriyan, J. c-Src binds to the cancer drug imatinib with an inactive Abl/c-Kit conformation and a distributed thermodynamic penalty. *Structure* **2007**, *15*, 299–311.

(19) 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.

(20) Belongie, S.; Malik, J.; Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 509–522.

(21) Le Cun, Y.; Cortes, C. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (accessed Feb 1, 2010).

(22) Richmond, N. J.; Willett, P.; Clark, R. D. Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graphics Modell.* **2004**, *23*, 199–209.

(23) Rohlf, F. J.; Slice, D. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* **1990**, *39*, 40–59.

(24) *Flap*, version May 2009; Molecular Discovery LTD: London, U.K. Molecular Discovery. http://www.moldiscovery.com (accessed July 1, 2009).

(25) Jonker, R.; Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **1987**, *38*, 325–340.

(26) Hicklin, J.; Moler, C.; Webb, P.; Boisvert, R.; Miller, B.; Pozo, R.; Remington, K. Jama: a Java matrix package. http://math.nist.gov/javanumerics/jama (accessed Nov 1, 2009).

(27) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365–370.

(28) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201–212.

(29) Oikonomakos, N. G.; Schnier, J. B.; Zographos, S. E.; Skamnaki, V. T.; Tsitsanou, K. E.; Johnson, L. N. Flavopiridol Inhibits Glycogen Phosphorylase by Binding at the Inhibitor Site. *J. Biol. Chem.* **2000**, *275*, 34566–34573.

(30) Deininger, M. W.; Druker, B. J. Specific targeted therapy of chronic myelogenous leukemia with imatinib. *Pharmacol. Rev.* **2003**, *55*, 401–423.

(31) Kumar, R.; Crouthamel, M. C.; Rominger, D. H.; Gontarek, R. R.; Tummino, P. J.; Levin, R. A.; King, A. G. Myelosuppression and kinase selectivity of multikinase angiogenesis inhibitors. *Br. J. Cancer* **2009**, *101*, 1717–1723.

(32) Vulpetti, A.; Bosotti, R. Sequence and structural analysis of kinase ATP pocket residues. *Il Farmaco* **2004**, *59*, 759–765.

(33) Tanaka, M.; Sagawa, S.; Hoshi, J.; Shimoma, F.; Matsuda, I.; Sakoda, K.; Sasase, T.; Shindo, M.; Inaba, T. Synthesis of anilino- monoindolylmaleimides as potent and selective PKCb inhibitors. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5171–5174.

(34) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biot.* **2007**, *25*, 197–206.

(35) *Pipeline Pilot*, version 7.5; Accelrys: San Diego, CA, 2008. Accelrys Pipeline Pilot http://accelrys.com/products/pipeline-pilot/ (accessed July 1, 2009).

(36) Higgins, D. G.; Bleasby, A. J.; Fuchs, R.; CLUSTAL, V. improved software for multiple sequence alignment. *Comp. Appl. Biosci.* **1992**, *8*, 189–191.

(37) Corbin, J. D.; Francis, S. H. Pharmacology of phosphodiesterase-5 inhibitors. *Int. J. Clin. Pract.* **2002**, *56*, 453–459.

(38) Sitruk-Ware, R. New progestagens for contraceptive use. *Human Reprod. Update* **2006**, *12*, 169–178.

CI1001263