Predicting Drug Promiscuity Using Spherical Harmonic Surface Shape-Based Similarity Comparisons

Violeta I. Pérez-Nueno*, Vishwesh Venkatraman, Lazaros Mavridis and David W. Ritchie*

INRIA Nancy – Grand Est, 615 rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France

Abstract: Polypharmacology is becoming an increasingly important aspect in drug design. Pharmaceutical companies are discovering more and more cases in which multiple drugs bind to a given target (promiscuous targets) and in which a given drug binds to more than one target (promiscuous ligands). These phenomena are clearly of great importance when considering drug side-effects. In the last 4 years, more than 30 drugs have been tested against more than 40 novel secondary targets based on promiscuity predictions. Current methods for predicting promiscuity typically aim to relate protein receptors according to their primary sequences, the similarity of their ligands, and more recently, the similarity of their ligand binding pockets.

Here, we present a spherical harmonic (SH) surface shape-based approach to predict rapidly promiscuous ligands and targets by comparing sets of SH ligand and protein shapes, respectively. We present details of our approach applied to a wide range of PDB complexes comprising ligands in a selected subset of the *MDL Drug Data Report* (MDDR) database which are distributed over 249 diverse pharmacological targets. The shape similarity of each ligand to each target's ligand set is quantified and used to predict promiscuity. We also analyse the correlation between binding pocket and ligand shapes. We compare our promiscuity predictions with experimental activity values extracted from the BindingDB database.

INTRODUCTION

Drug promiscuity may be defined as the specific binding of a drug-like molecule to more than one target. On the other hand, if a protein binds different ligands, it can be considered as a promiscuous receptor [1]. These are notions illustrated in Fig. (**1**). The concept of 'target-hopping,' whereby a binder for one target can be considered as the basis for leads for another target has historically been extremely fruitful in lead discovery [2]. Nowadays, polypharmacology is becoming an increasingly important aspect in drug design. In the last 4 years, more than 30 drugs have been tested against more than 40 novel secondary targets based on promiscuity predictions [3]. Pharmaceutical companies are discovering more and more cases in which multiple drugs bind to a given target (promiscuous targets) and in which a given drug binds to more than one target (promiscuous ligands). Both of these phenomena are clearly of great importance when considering drug side-effects. For example, a common reason for terminating a drug development program is that the leads are found to be non-selective or promiscuous [4]. Thus, the *in silico* prediction of unwanted side effects caused by the promiscuous behaviour of drugs and their targets is highly relevant to the pharmaceutical industry. Considerable effort is now being put into the computational [5, 6] and experimental [7, 8] screening of several suspected off-target proteins in the hope that side effects might be identified

early, before the cost associated with developing a drug candidate rises steeply [9]. On the other hand, promiscuity is not always unwelcome and it can even be exploited for drug development. The use of old drugs for new targets has been shown to provide a promising way to reduce both the time and cost of drug development [10].

Given that it is currently infeasible to screen a drug against all of the proteins expressed by the human genome, several computational techniques have been developed to predict the pharmacological profiles of known drugs. These range from the well-known docking of compounds into protein structures to the use of machine learning methods [11, 12], sequence comparison [13] side-effect similarity [6], and fingerprint/pharmacophore comparison methods [14, 15]. These *in silico* methods typically aim to relate protein receptors to each other quantitatively based on their similarity in primary sequence space [14], ligand chemical descriptor space [15], and more recently in their pharmacophoric pocket descriptor space [16].

The first and most common techniques use *e.g.* the BLAST [17, 18] or FASTA [19] sequence alignment tools to create similarity maps in protein sequence space [20]. Later, other approaches were developed which compare the chemistry of the targets' ligands following the hypothesis that similar molecules are likely to have similar properties [21]. For example, Keiser *et al*., relate receptors to each other according to the chemical similarity of their ligands. In their Similarity Ensemble Approach (SEA) [22], the calculated probability that two molecules might interact with the same target by chance is expressed using an expectation

*Address correspondence to these authors at the INRIA Nancy – Grand Est, 615 rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France; Tel: + +33-3-83593045; Fax: + +33-3-83413079; E-mails: violeta.pereznueno@inria.fr, dave.ritchie@inria.fr

value [23], which is conceptually similar to the E-value used in sequence alignment software such as BLAST [17]. Mestres *et al.,* relate proteins in ligand space using three in–house molecular descriptors (PHRAG, FPD, SHED) [24, 25]. With the increase in recent years of approaches for comparing protein pockets [26] such as four-point pharmacophoric descriptors (FLAP [27]), three-point pharmacophoric descriptors (Cavbase [28], SiteEngine [29], SuMo [30]), geometric hashing methods (Kinnings and Jackson [31]), and graph-matching-base algorithms (IsoCleft [32]), the notion of a *binding pocket similarity space* has recently been proposed. This is based on the principal that protein binding pockets are the place where the interactions between a protein and a ligand are formed. Hence, they must have complementary shapes and physicochemical properties to the small molecules that they accommodate. Therefore, calculating the similarity of binding pockets allows proteins with similar function and selectivity for the same binding partners to be related. For example, Weskamp *et al.* compared targets by the similarity of their binding pockets using the LIGSITE program [14]. Milleti *et al.*, recently related receptors to each other using pocket-based "shape context" descriptors [16, 33, 34].

Here, we present a shape-based approach which uses spherical harmonic (SH) representations [35, 36] to compare molecular surfaces efficiently. This approach relates receptors to each other by the SH shape similarity of their ligands and their binding pockets. Since shape complementarity is an essential feature for molecular recognition, using ligand and binding pocket shapes should provide a good way to characterise their properties. If two binding pockets of different proteins share a common shape, it is likely that ligands that bind to part of one binding pocket will also be recognized in the corresponding part of the other pocket. On the other hand, if two ligands of different proteins share a similar shape, it is likely that both of them will complement the shape of each binding pocket. Hence, by identifying similar ligands and binding pocket shapes, our approach aims to provide a shape-based way to predict promiscuous ligands and targets.

We present the results of applying our approach to a wide range of ligands which Shuffenhauer *et al.,* [37] previously selected from the MDL Drug Data Report (MDDR) database [38], and for which crystallographic protein-ligand complexes exist in the Protein Data Bank (PDB) [39]. This gives an annotated list of ligands for 249 protein targets of pharmacological interest. The shape similarity between ligands and between binding pockets for these selected protein targets is quantified and used to predict promiscuity. We analyse the correlation between binding pocket and ligand shape spaces. We also compare our promiscuity predictions with experimental activity values extracted from the BindingDB database [40].

METHODS

Calculating SH Shapes

We use the PARASURF and PARAFIT modules [41] to calculate and superpose SH molecular surfaces, and the MSSH [42] program to calculate the SH shapes of protein pockets. PARASURF calculates molecular shape and electronic properties from semi-empirical quantum

mechanics theory and encodes these properties as SH expansions [35, 36]. Surface shapes are represented as radial distance expansions of the molecular surface, $r(\theta,\phi)$, with respect to a selected harmonic coordinate origin (CoH), which is normally set equal to the molecular center of gravity (CoG). This allows an entire molecular surface shape to be captured using a Fourier-like polynomial expansion, such as Equation 1

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi)$$ Equation 1

where θ and ϕ are the spherical coordinates, $y_{lm}(\theta,\varphi)$ are real spherical harmonics, a_{lm} are the expansion coefficients, and L is the order or highest polynomial power of the expansion. As determined previously, we use $L = 6$ for shape comparisons [43]. Mathematically, the SH approach applies only to "starlike" shapes which have single-valued surfaces with respect to radial rays projections from the chosen CoH. Most molecules do not satisfy this requirement. Hence, the SH surfaces described here should be considered as "surface envelopes" which enclose the true molecular surface. For highly non-starlike molecules this surface envelope can be a rather severe approximation to the true surface. Indeed, in extreme cases, the CoH can fall outside the molecular surface, and this can cause the software to fail.

PARAFIT and MSSH calculate shape superpositions by exploiting the special rotational properties of the SH functions. For example, rotated SH expansion coefficients for a molecule B can be calculated as

$$b'_{lm} = \sum_{m=-l}^{l} R^{(l)}_{mm'}(\alpha,\beta,\gamma) b_{lm}$$ Equation 2

where (α,β,γ) are zyz Euler rotation angles and $R^{(l)}(\alpha,\beta,\gamma)$ is the *l'th* real Wigner rotation matrix. To calculate a superposition between a pair of molecules A and B, the harmonic coordinate origin (CoH) of molecule B is translated to that of the fixed reference molecule A, and a rotational search to a resolution of 1 degree in each Euler rotation angle is then performed to find the rotation which minimizes the distance, D_{AB}, between the corresponding pairs of SH expansions (Equation 3).

$$D_{AB} = \int \left(r_A(\theta,\phi) - r'_B(\theta,\phi) \right)^2 d\Omega$$ Equation 3

Thanks to the orthogonality of the basis functions, Equation 3 reduces to

$$D_{AB} = |a|^2 + |b|^2 - 2ab$$ Equation 4

where a and b represents vectors of expansion coefficients.

In practice, it is often more convenient to compare molecules using a normalised similarity score such as Equation 5 [43, 44].

$$S = \frac{ab}{\left(|a|^2 + |b|^2 - ab\right)}$$ Equation 5

Using PARAFIT it is also straight-forward to construct the average or "consensus" shape, $\bar{r}(\theta,\phi)$, of a group of N molecules by calculating the average of their superposed SH expansion coefficients, a^k_{lm}, for $k=1,..., N$, as shown in Equation 6 [44].
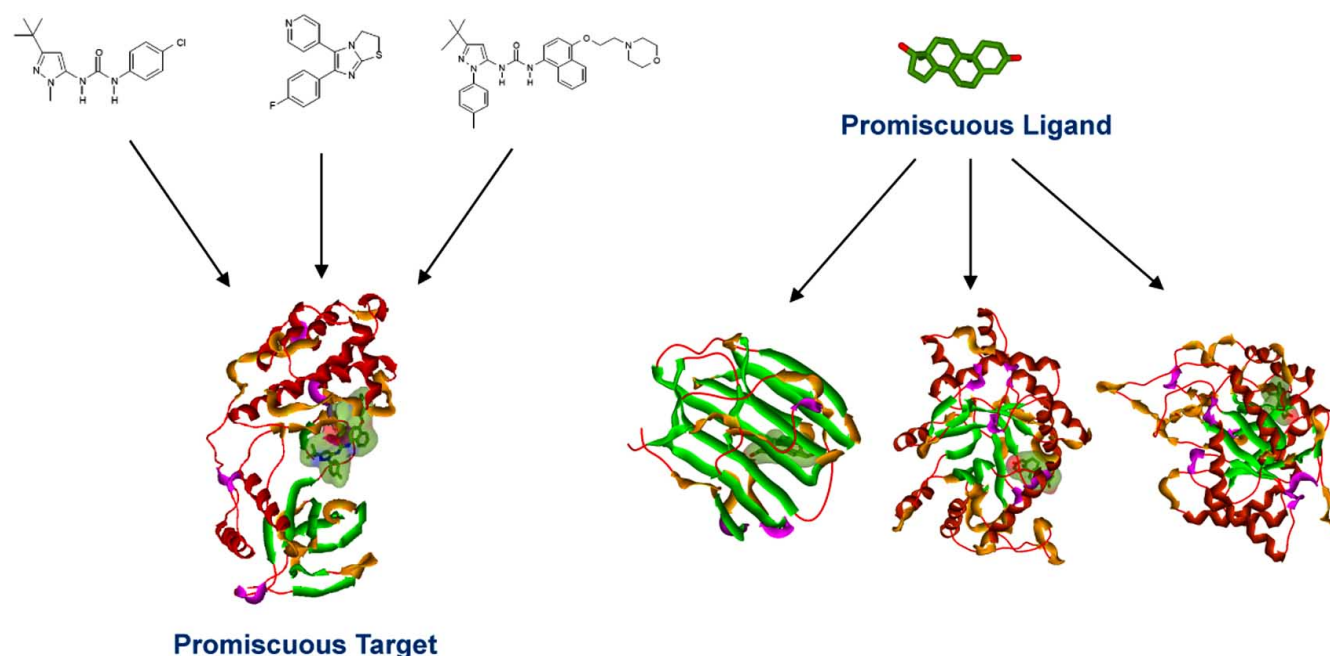
Fig. (1). Drug Promiscuity. Left: multiple drugs bind to a given target (promiscuous target). Right: a given drug binds to more than one target (promiscuous ligand).

$$\bar{r}(\theta,\phi) = \frac{1}{N}\sum_{k=1}^{N}\sum_{l=0}^{L}\sum_{m=-l}^{l} a_{lm}^{k} y_{lm}(\theta,\phi)$$ Equation 6

However, before computing the average, each molecule in the consensus must first be rotated to minimize the distance between it and the remaining N - 1 molecules. In practice, because these rotations are not known a priori, the consensus shape is constructed iteratively as described previously [44]. We have shown that the consensus shape-based representation can be used to capture the essential 3D shape features of several known high-affinity ligands and to encode them in the form of a single representative pseudo-molecule which can be used as a VS query [45-48].

The SH consensus approach may also be used to calculate the consensus shapes of both ligand molecules and protein receptor pockets. Here, the SH coefficients of the binding pockets are computed using MSSH. Each pocket surface is calculated around the bound ligand coordinates using the default MSSH radial cut-off distance of 20 Å. It is worth noting that MSSH can usually calculate a good representation of closed pockets, but open pockets on the protein surface are not represented well, and this limits the quality of the results. Nonetheless, all pockets are analysed. Once the SH coefficients of the binding pockets are calculated, the same consensus algorithm is applied as described above.

Ligand and Target Data Preparation

We applied our approach to those PDB complexes which contain the ligands in Schuffenhauer's subset of the MDDR database (version 2010.2) [38], comprising 65367 compounds distributed over 249 diverse pharmacological targets for which experimental binding information is known. In other words, we consider only crystallographically determined ligand coordinates and we do not explicitly consider multiple ligand conformations. On the other hand, calculating SH consensus shapes can take into account 3D shape features of several conformations to be used as an averaged pseudo-molecule [48]. Schuffenhauer's subset is based on a simple ontology [37] that maps Enzyme Commission (EC) numbers [49], GPCRs, ion channels, and nuclear receptors to MDDR activity classes. In this dataset, all ligands are annotated where possible according to the MDDR activity class. It should be noted that these ligand annotations can be rather general text descriptions, such as "dihydrofolate reductase inhibitor" or "anticancer agent", for example. Therefore, an annotation can correspond to multiple targets. Supplementary Table **1** shows the list of target annotations used for our promiscuity predictions.

Fig. (**2**) shows the workflow followed to process the data for our promiscuity predictions. The MDDR database was filtered according to Schuffenhauer's subset to give 8659 ligands with 196 unique annotations. These ligands were then used to search the PDB hetero-atom dictionary to find ligands for which crystal structures exist. These structures were extracted from the PDB using their three-letter PDB ligand codes to give a total of 957 protein-ligand complexes. Any structures solved by NMR and those without a CATH code were removed to give 687 proteins belonging to 76 unique annotations. Only one chain was kept for each protein.

Predicting Promiscuity Using SH Shape-Based Similarity

Ligand SH Shape Similarity

To predict ligand promiscuity, all the ligands from the PDB complexes were extracted and transformed into a canonical orientation using PARAFIT. The SH similarity of each ligand with each target's ligand set was calculated to give an all-*vs*-all ligand interaction matrix. The matrix was then analyzed by using three Tanimoto thresholds.

Column 1 (left flow):

MDDR 2D SDF: 309 targets
→ Remove duplicates

SHUFFENHAUER DATA: 213 ANNOTATIONS
→ 17 non Activity key correspondences between Schuffenhauer version and new MDDR version

8659 UNIQUE LIGANDS distributed in 196 UNIQUE ANNOTATIONS
→ Search in the PDB Heteroatom dictionary the corresponding ligand 3-letter code names for the list of ligands in each Schuffenhauer annotation

424 UNIQUE LIGANDS corresponding to 93 UNIQUE ANNOTATIONS involving at least one ligand found in the PDB Heteroatom dictionary
→ Search the PDB to retrieve the pdb files associated with those 3-letter code ligands

957 PDB codes
→ Remove NMR structures and structures without CATH code

687 PDB codes (kept one chain for each PDB) belonging to 76 UNIQUE ANNOTATIONS

Column 2 (middle flow):

LIGANDS
→ Extract all the ligands from the 687 PDBs

687 LIGANDS
→ Superpose all ligands onto the origin (the canonical orientation) with PARASURF/PARAFIT

677 LIGANDS
→ 10 compounds fall when calculating PARASURF (CoG outside molecule). Calculate ligand SH shape-similarity with PARAFIT

ALL VS ALL LIGAND INTERACTION MATRIX

Column 3 (right flow):

BINDING POCKETS
→ Align complexes into the ligand canonical orientation

677 BINDING POCKETS
→ Calculate binding pocket SH shape with MSSH. 10 compounds fail when calculating MSSH

670 BINDING POCKETS
→ Consensus of the pockets belonging to the same annotation

76 CONSENSUS POCKETS
→ Calculate binding pocket SH shape-similarity of the consensus pockets

ALL VS ALL BINDING POCKET INTERACTION MATRIX

(converging):

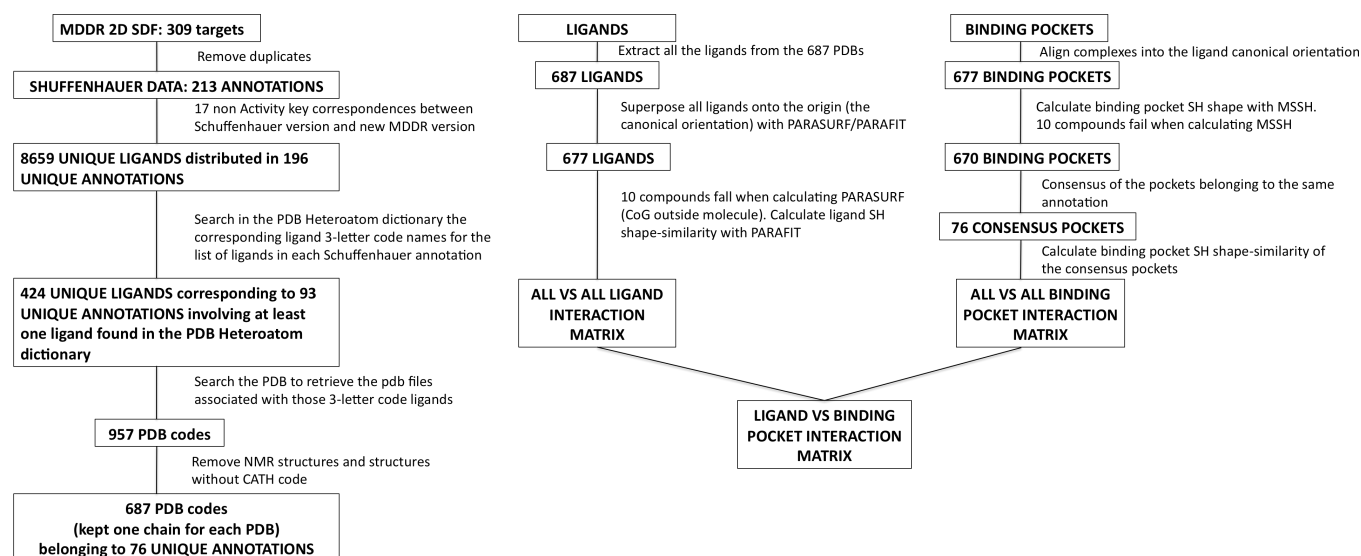LIGAND VS BINDING POCKET INTERACTION MATRIX

Fig. (2). Data processing flowchart. We used a subset of the MDDR database comprising 76 diverse pharmacological targets having at least one ligand in the PDB heteroatom dictionary. The similarity of each ligand SH shape to each target's ligand set shape is calculated and quantified using the Tanimoto coefficient. The similarity between each SH consensus shape pocket to each target's consensus binding pocket is calculated in the same way. Finally, the ligand-pocket shape interaction matrix is analyzed to predict promiscuity.

Binding Pocket SH Shape Similarity

For promiscuity prediction in binding pocket space, we positioned all protein-ligand complexes according to the canonical orientations of their ligands [50]. This placed all pockets in a ligand-defined standard orientation in order to calculate SH binding pocket shapes using MSSH. The consensus shapes of the pockets belonging to the same MDDR annotation were calculated with PARAFIT. The SH similarity between consensus pocket shapes was calculated to give an all-*vs*-all binding pocket interaction matrix. As before, this matrix was analyzed using the same three Tanimoto thresholds.

Ligand *vs* Binding Pocket SH Shape Similarity

Finally, a ligand-pocket SH shape interaction matrix was also analyzed in the same way.

RESULTS

Ligand-Ligand Interaction Matrix

Fig. (**3a**) shows the form of the 677 ligand-ligand interaction matrix. A high resolution zoom of a portion of this figure is shown in Fig. (**4**). A ligand Tanimoto score greater than 0.9 is shown in dark blue. Ligands with a Tanimoto score between 0.7 and 0.9 are shown in blue, and scores lower than 0.7 are shown in light blue. Both axes are labelled according to the MDDR annotations to which the 677 ligands belong. It can be observed that each ligand shape matches itself (diagonal in dark blue). Similarly, ligands belonging to related targets are found to have similar shapes (dark blue areas) such as some nuclear hormone receptors (vitamin d3-like receptors, estrogen, androgen, progesterone) and serine proteases (coagulation factors Xa and VIIIa, thrombin, trypsin, β-lactamase).

Pocket – Pocket Interaction Matrix

Fig. (**3b**) (high resolution in Fig. **5**) shows the 76 consensus pocket-pocket interaction matrix. It is color coded and labelled as in Fig. (**3a**). Each annotation is represented by a consensus binding pocket shape. The interaction matrix is clustered by consensus binding pocket shape similarity. It can be seen that this matrix also groups together related targets. For example, several distinct groups each with high shape similarity (Tanimoto $\geq$ 0.9) are found for the serine proteases, other proteases, nuclear hormone receptors, kinases, GPCRs, ion channels, enzymes and metallo enzymes. More specifically, the serine proteases cluster includes coagulation factors Xa and VIIIa, thrombin, trypsin, β-lactamase, interleukin-8, and serine type-d ala-d-ala carboxypeptidase. The main nuclear hormone receptor cluster includes the vitamin d3-like receptors, estrogen, and androgen. Similarly, the GPCR cluster includes the β-adrenoreceptor type 1, β-adrenoreceptor, ETA, and endothelin. The main enzyme cluster includes the peptidase, serine endopeptidase, alcohol dehydrogenase, purine nucleoside phosphorilase hydroxymethylglutaryl CoA reductase, cholestenone 5α-reductase, adenosylhomocysteinase, lanosterol synthase, and guanylate cyclase. It can also be seen that each consensus binding pocket matches itself (diagonal in dark blue). Hence, it can be seen that comparing receptor pocket shapes correctly groups many protein targets into their expected macromolecular target family.

Analysis of the pocket-pocket matrix shows some highly similar binding pockets. Fig. (**6**) shows four examples of shape supperpositions between the correlated pockets. Firstly, the consensus pocket shape of androgen (Fig. **6** top) shares high shape similarity with those of β-adrenoreceptor, purine nucleoside phosphorilase, thimidine kinase, hydroxy-methylglutaryl CoA reductase, and adenosylhomocysteinase. The promiscuity predicted for androgen is consistent with several existing MDDR activity classes (*e.g.* androgen, aromatase inhibitor, antiandrogen) for the androgen ligands. Secondly, the consensus pocket of hydroxymethylglutaryl CoA reductase (Fig. **6** middle) shares high shape similarity with the consensus binding pockets of lanosterol synthase,
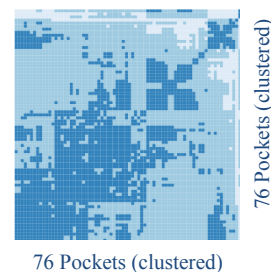
a) Ligand *vs* ligand

Tanimoto ≥ 0.9 *Dark blue*
0.7 ≤ Tanimoto < 0.9 *Blue*
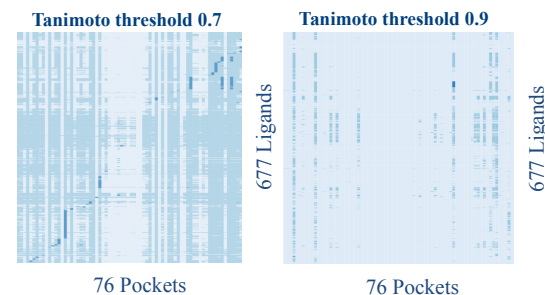   Tanimoto < 0.7 *Light blue*
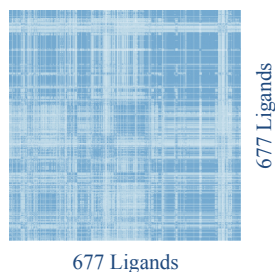
b) Pocket *vs* pocket

Tanimoto ≥ 0.9 *Dark blue*
0.7 ≤ Tanimoto < 0.9 *Blue*
            Tanimoto < 0.7 *Light blue*

c) Ligand *vs* pocket

Similarity found & PDB complex exists *Dark blue*
Similarity found & no PDB complex exists *Blue*
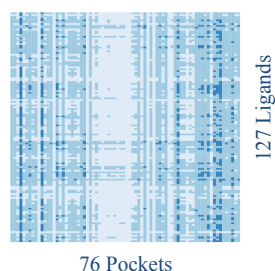No similarity found *Light blue*

Tanimoto threshold 0.7   **Tanimoto threshold 0.9**

677 Ligands   76 Pockets (clustered)   677 Ligands   677 Ligands

677 Ligands   76 Pockets (clustered)   76 Pockets   76 Pockets

d) *In silico* vs *in vitro* interaction matrices

In silico

Tanimoto ≥ 0.9 *Dark blue*
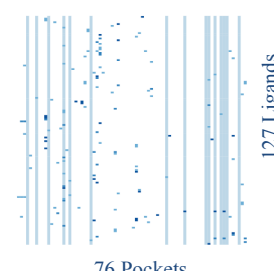0.7 ≤ Tanimoto < 0.9 *Blue*
            Tanimoto < 0.7 *Light blue*

In vitro (IC50/nM)

0 < IC50/nM ≤ 1 *Dark blue*
1 < IC50/nM ≤ 10 *Blue*
            IC50/nM > 10 *Light blue*

In vitro vs in silico

127 Ligands   127 Ligands   127 Ligands
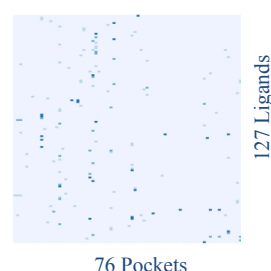
76 Pockets   76 Pockets   76 Pockets

Fig. (3). Predicting Promiscuity Using SH Shape-Based Similarity. This figure shows illustrative images of **a**) the ligand-ligand interaction matrix, **b**) the binding pocket-binding pocket interaction matrix, **c**) the ligand-binding pocket interaction matrix, **d**) the *in silico* vs *in vitro* interaction matrices.

purine nucleotide phosphorilase, thimidine kinase, cholestenone 5α-reductase, and adenosylhomocysteinase. Several MDDR activity classes (*e.g.* hypolipidemic, HMG-CoA reductase β inhibitor) are also related to the hydroxymethylglutaryl CoA reductase ligands. Thirdly, GABA-A alpha is predicted to be another highly promiscuous target. Its pocket shows high shape similarity with the pockets of β-adrenoreceptor, purine nucleoside phosphorilase, thimidine kinase, hydroxymethylglutaryl CoA reductase, cholestenone 5α-reductase, adenosylhomo-cysteinase, lanosterolsynthase, vitamin d3-like receptors, estrogen, androgen, adenosindeaminase, acetylcholine-sterase, and RNA directed DNA polymerase. This predicted high promiscuity is also supported by the large number of related MDDR activity classes for this annotation (non opioid analgesic, GABA-A/benzodiazepine receptor, sedative/hypnotic, anxiolytic, agent for sleep disorders, benzodiazepine agonist, alcohol deterrent, anticonvulsant, agent for premedication, antimigraine, and intravenous anesthetic). Finally, the thrombin consensus binding pocket is found to be similar to the consensus pocket shapes of coagulation factors Xa and VIIIa, β-lactamase, trypsin, interleukin-8, and serine-type d-ala-d-ala carboxypeptidase. This also agrees with the activity classes of the thrombin ligands in MDDR (*e.g.* anticoagulant, thrombin inhibitor, factox Xa inhibitor, trypsin inhibitor, protease inhibitor).

Overall, analysis of the pocket interaction matrix in Fig. **5** points to the prediction of several very promiscuous targets (dark blue rows) such as the aforementioned examples, *i.e.* GABA-A alpha subunit, androgen, hydroxymethylglutaryl CoA reductase, and thrombin, as well as estrogen, vitamin 3d-like, acetylcholinesterase, RNA directed DNA polymerase, and purine nucleoside phosphorylase, and also selective targets such as caspase (light blue row). These predictions agree with several related activity classes found in the MDDR for each of these annotations.

Ligand-Pocket Interaction Matrix

Fig. (**3c**) shows the ligand-consensus binding pocket interaction matrix using a Tanimoto threshold of 0.7 and a Tanimoto threshold of 0.9. A high resolution zoom of portions of these figures is shown in Figs. (**7** and **8**), respectively. The x-axis represents the 76 annotation consensus binding pockets and the y-axis the 677 ligands distributed in those annotations. Both axes are labelled according to the MDDR annotations to which the 677 complexes belong. The matrix with the lower threshold highlights the confirmed interactions according to the crystallised structures present in the PDB. It can be observed that each ligand shape matches the annotation-based consensus pocket that the ligand belongs to (diagonal in dark blue). A lower and more permissive threshold helps to

Fig. (4). High resolution zoom of the top right corner (red box) of the ligand-ligand SH shape interaction matrix. The interaction matrix is color coded using darker blue for a high shape similarity (Tanimoto ≥ 0.9), blue for medium similarity (Tanimoto between 0.7 and 0.9) and light blue for low similarity (Tanimoto < 0.7). Both axes are labelled according to the MDDR annotations in which the ligands are distributed.

identify the similarities between known ligands for a given target and some other ligands but it causes quite a lot of false positives (large blue areas), whereas using a more restrictive threshold clearly highlights the predicted promiscuous targets (*i.e.* a high number of dark blue pixels in a single column). The 0.9 threshold matrix highlights as blue columns several predicted promiscuous targets which were identified previously in the ligand-ligand and pocket-pocket matrices (*e.g.* GABA-A alpha subunit, androgen, or hydroxymethylglutaryl CoA reductase) as well as some others (procollagen proline dioxygenase, cholestenone 5α

reductase, glutamate receptor kainite subunit, estrogen, vitamin d3-like receptors, lanosterol synthase, adenosylhomocysteinase, acetylcholinesterase, RNA directed DNA polymerase, thymidine kinase, purine nucleoside phosphorylase and ribonucleoside diphosphate reductase). The fact that there exist in the PDB several complexes with ligands for different targets confirms these predictions (dark blue columns highlighted in the 0.7 threshold matrix). It is interesting to note that the caspase pocket (which appears in the centre of Fig. **7**, **8**, and **9**) has a completely light blue column indicating that this is a highly specific target.

Fig. (5). High resolution pocket-pocket SH shape interaction matrix. The interaction matrix is color coded using darker blue for a high shape similarity (Tanimoto ≥ 0.9), blue for medium similarity (Tanimoto between 0.7 and 0.9) and light blue for low similarity (Tanimoto < 0.7). Both axes are labelled according to the 76 annotations in which the 677 complexes are distributed. Each annotation is represented by a consensus binding pocket shape. The interaction matrix is clustered by consensus binding pocket shape similarity. This also groups together related targets.

This analysis of the ligand-pocket matrix mainly agrees with the correlations found between binding pockets. Fig. (**10**) shows the ligand-pocket superpositions of the same consensus pockets shown in Fig. (**6**) with selected examples of high similarity ligands. For example, the androgen consensus binding pocket (Fig. **10** top) shares high shape similarity with the ligands of β-adrenoreceptor, purine nucleoside phosphorilase, thimidine kinase, hydroxymethyl-glutaryl CoA reductase, and adenosylhomocysteinase. Similarly, the hydroxymethylglutaryl CoA reductase consensus pocket (Fig. **10** middle) shares high shape similarity with the ligands of lanosterol synthase, purine nucleotide phosphorilase, thimidine kinase, cholestenone 5α-reductase, and adenosylhomocysteinase.

Unlike the pocket-pocket matrix, the thrombin consensus pocket is found to have lower similarity with other serine protease ligands. The pocket-ligand superpositions look worse in this case. This is because MSSH does not represent accurately the shapes of surface pockets. Fig. (**10** bottom)
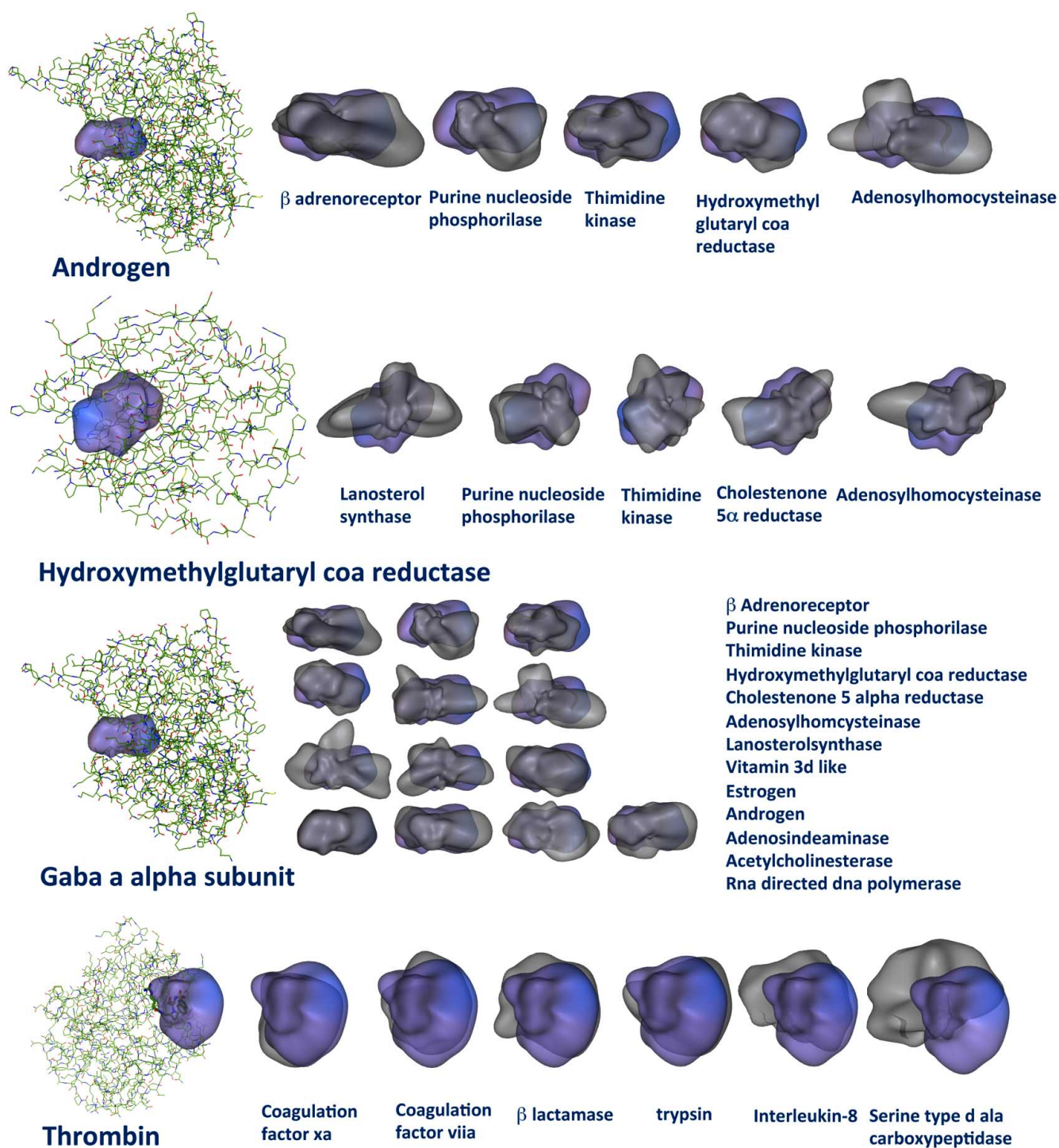
Fig. (6). Example superpositions of consensus pocket shapes. This figure shows the consensus pocket shapes of four example targets (androgen, hydroxymethilglutaryl CoA reductase, GABA-A alpha subunit, and thrombin) in lilac, along with the superpositions of the pockets of several other targets with similar consensus pocket shapes (grey).

shows the thrombin consensus pocket superposed with the ligands of coagulation factors Xa and VIIIa, β-lactamase, trypsin, interleukin-8, and serine-type d-ala-d-ala carboxypeptidase. As can be seen, the thrombin SH pocket shape is bigger than the shapes of these ligands. Hence, only the large trypsin ligands match well this large surface pocket shape. Nevertheless, it can be observed in Fig. (**10**) that the

serine protease ligands superpose correctly on the left side of the thrombin pocket, while the poor quality surface pocket representation on the right remains unmatched. Finally, in the pocket-pocket correlation matrix, the GABA-A alpha is predicted to be a highly promiscuous target. The GABA-A alpha consensus pocket shows high shape similarity with the ligands of β-adrenoreceptor, purine nucleoside phosphori-
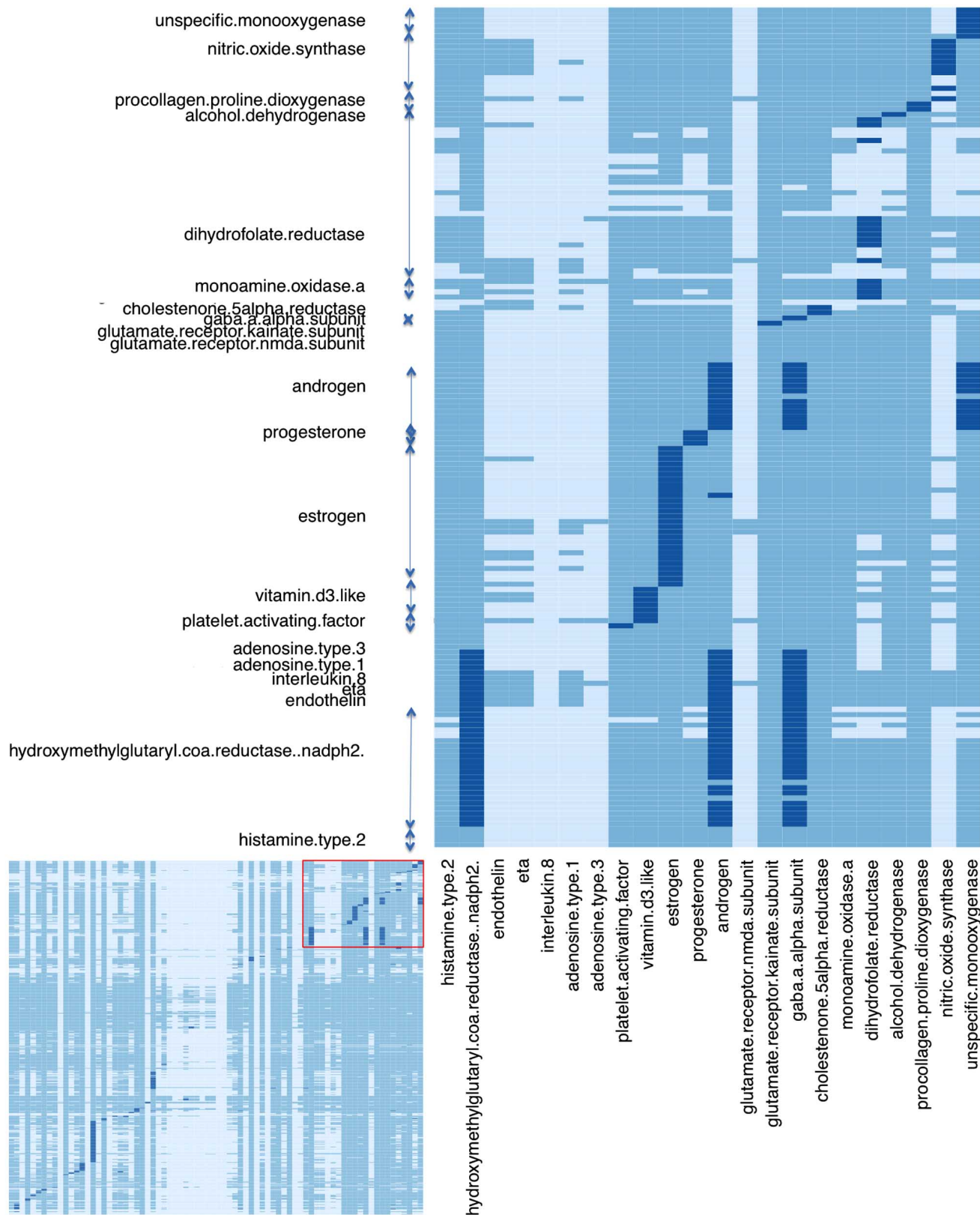
Fig. (7). High resolution zoom of the top right corner (red box) of the ligand-pocket SH shape interaction matrix at a Tanimoto threshold of 0.7. The interaction matrix is colored in dark blue if there is SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.7 and a PDB complex exists, blue if there is SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.7 and a PDB complex does not exist, and light blue if there is no SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.7. Both axes are labelled according to the MDDR annotations in which the complexes are distributed.
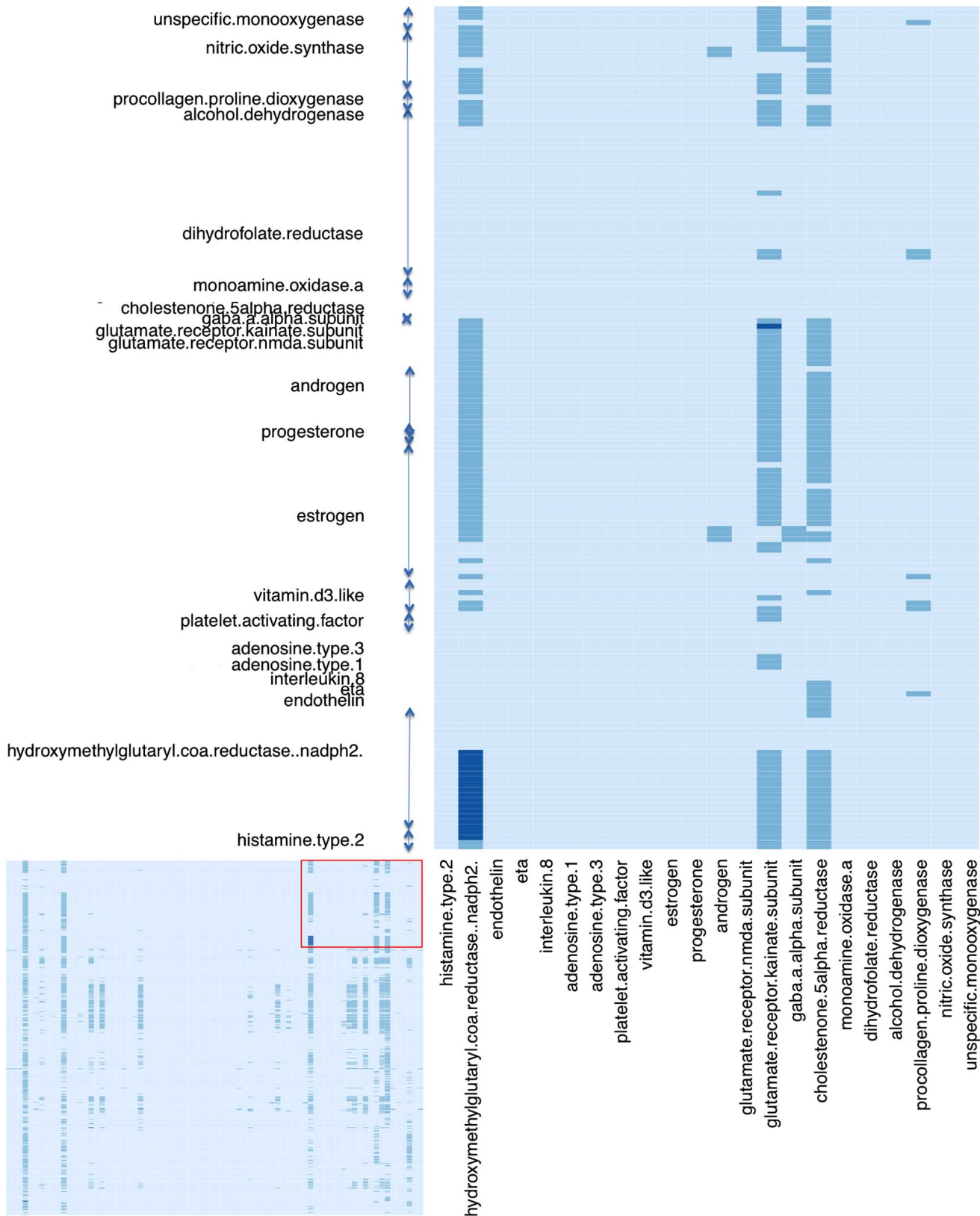
Fig. (8). High resolution zoom of the top right corner (red box) of the ligand-pocket SH shape interaction matrix at a Tanimoto threshold of 0.9. The interaction matrix is colored in dark blue if there is SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.9 and a PDB complex exists, blue if there is SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.9 and a PDB complex does not exist, and light blue if there is no SH similarity between ligand and binding pocket shapes over a Tanimoto of 0.9. Both axes are labelled according to the MDDR annotations in which the complexes are distributed. It can be observed that a high threshold value highlights targets which are predicted to be promiscuous (*i.e.* those columns with a high number of dark blue pixels).
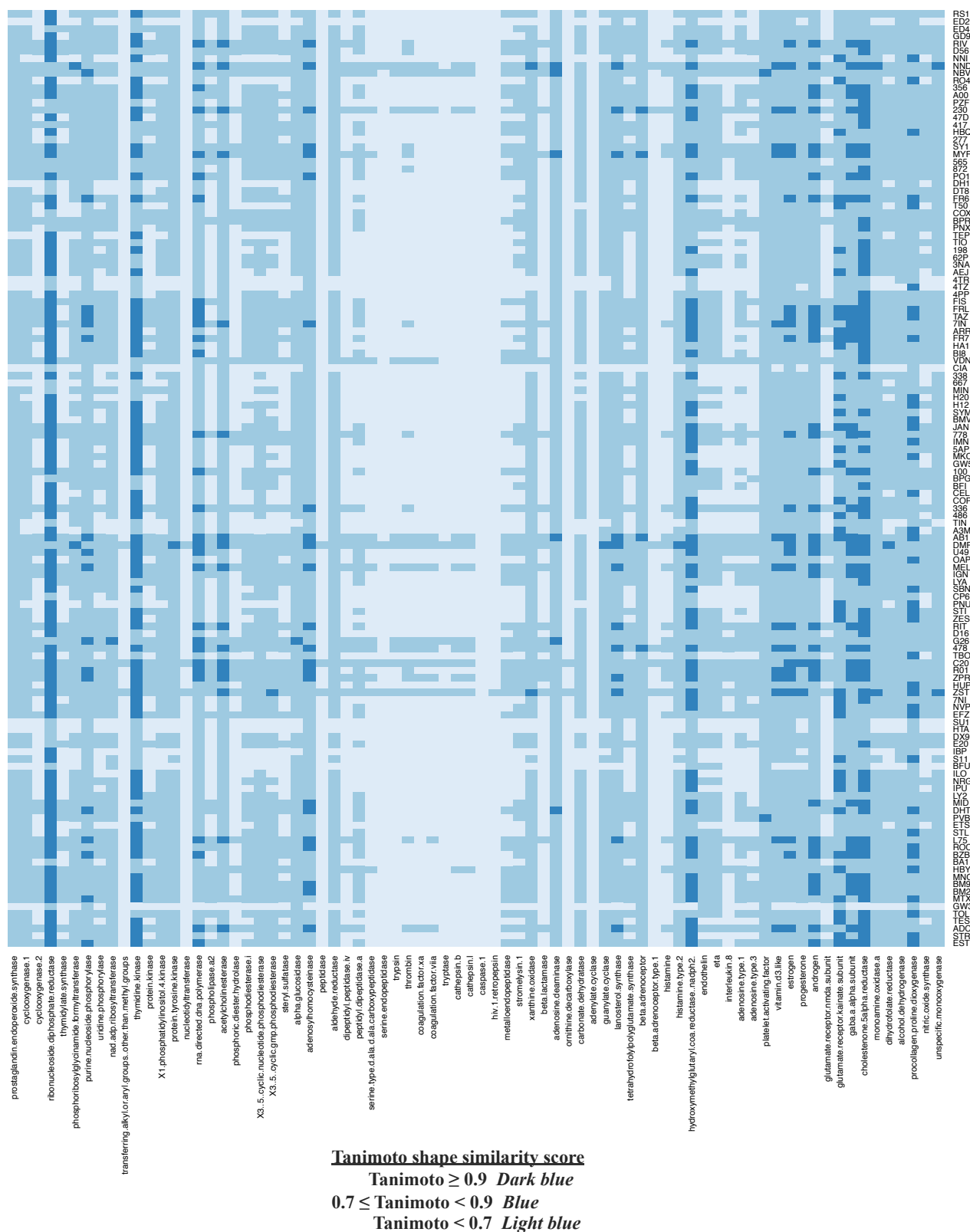
Fig. (9). High resolution *in silico* promiscuity prediction matrix. The *in silico* ligand-pocket interaction matrix is shown for the 127 ligands for which biological activities are available in BindingDB. The matrix is color coded using dark blue for a high shape similarity (Tanimoto ≥ 0.9), blue for medium similarity (Tanimoto between 0.7 and 0.9) and light blue for low similarity (Tanimoto < 0.7). The horizontal axis is labelled according to the 76 annotations in which the 677 complexes are distributed, and the vertical axis shows the 3-letter code for the ligands used in the promiscuity predictions. The targets with high shape similarity amongst all the ligands (*i.e.* columns with a high number of dark pixels) are predicted to be promiscuous.
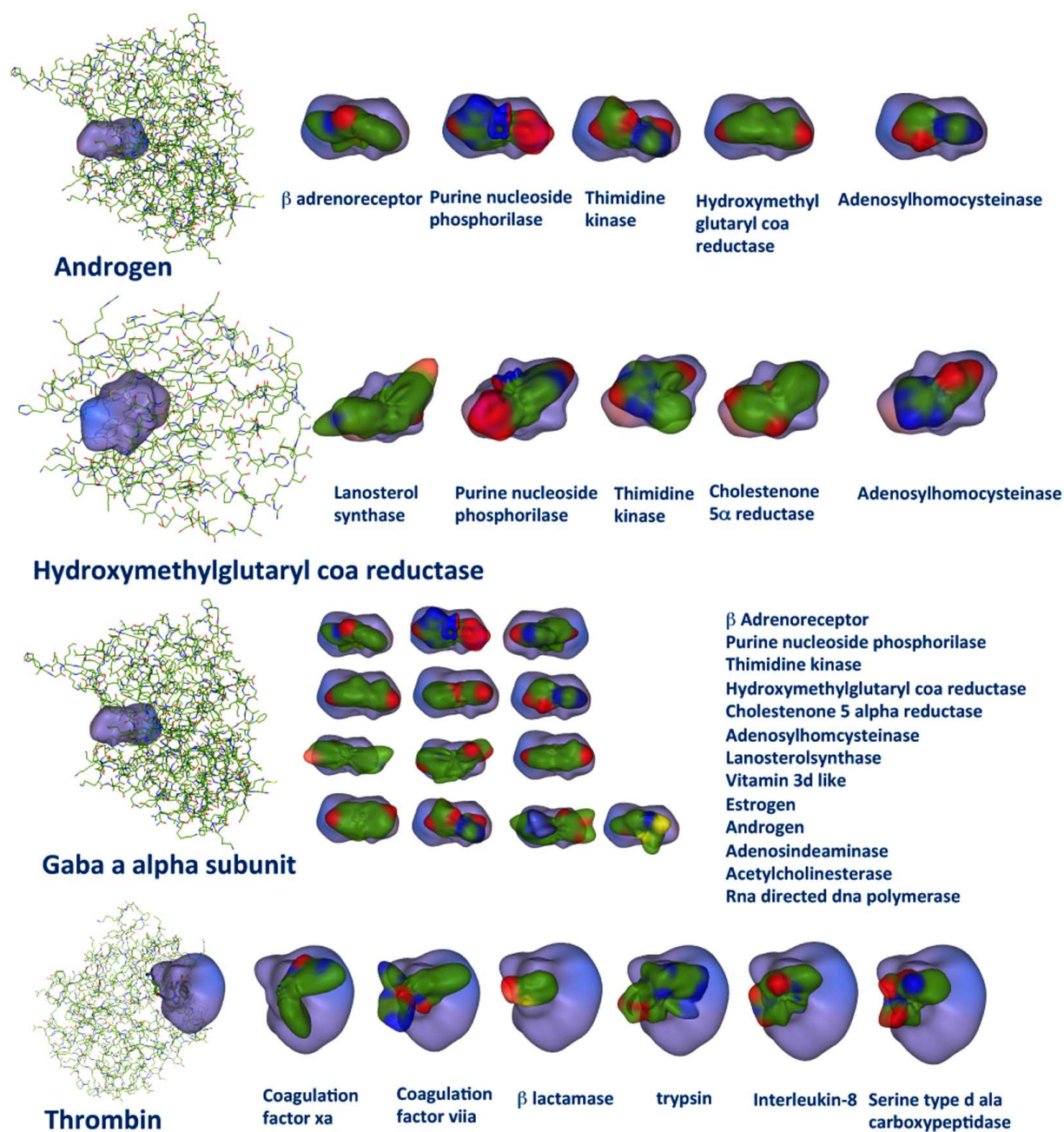
Fig. (10). Example superpositions of consensus pocket shapes with selected ligands. This figure shows the same consensus pockets as in Fig. (**6**) (lilac), with high similarity ligands shown as multi-coloured surfaces (red/green/blue/yellow).

lase, thimidine kinase, hydroxymethylglutaryl CoA reductase, cholestenone 5α-reductase, adenosylhomocysteinase, lanosterolsynthase, vitamin d3-like receptors, estrogen, androgen, adenosindeaminase, acetylcholinesterase, and RNA directed DNA polymerase. Again, the promiscuity predicted for all these targets agrees well with the existing MDDR activity classes for their ligands.

Comparison with Experimental Results

In order to validate our approach, we compared our *in silico* interaction matrices with biological activity data extracted from BindingDB [40]. Fig. (**3d**) compares our *in silico* results with the experimental results for the 127 ligands for which data is available in BindingDB. The *in*

silico ligand-pocket interaction matrix is color coded as in Fig. (**3a** and **3b**). Fig. (**9**) shows this *in silico* matrix in detail. As before, a high similarity threshold helps to highlight the possible promiscuous targets. Hence, we use a Tanimoto threshold of 0.9 to predict promiscuous targets (dark blue columns in Fig. **9**) and we compare these predictions with promiscuity evidence from *in vitro* results. Supplementary Table **2** lists the predicted promiscuous targets as well as several MDDR activity classes related to their ligands. The predicted promiscuous targets agree with the existing MDDR activity classes for their ligands except for vitamin d3-like receptors and thimidine kinase, which are only found to be related to vitamin 3D analog class and thymidine kinase inhibitor, respectively. Fig. (**11**) shows in detail the

$0 < \text{IC50/nM} \leq 1$ *Dark blue*
$1 < \text{IC50/nM} \leq 10$ *Blue*
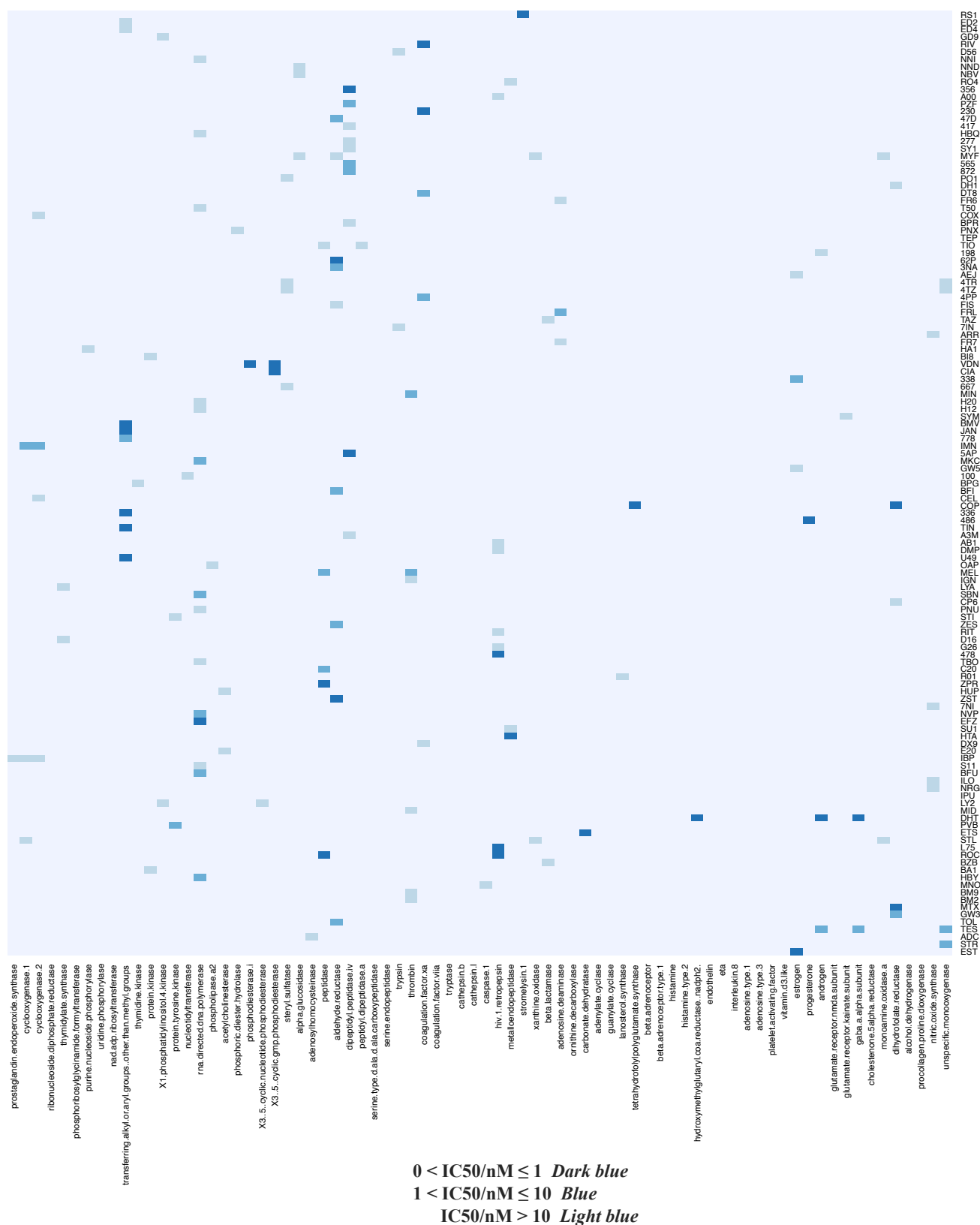$\text{IC50/nM} > 10$ *Light blue*

Fig. (11). High resolution *in vitro* promiscuity prediction matrix. The *in vitro* interaction matrix is color coded according to the biological activity values, using dark blue for high activity (IC50/nM ≤ 1), blue for medium activity (1 < IC50/nM ≤ 10) and light blue for low activity (IC50/nM > 10). The horizontal axis is labelled according to the 76 annotations in which the 677 complexes are distributed, and the vertical axis shows the 3-letter code for the ligands used in the promiscuity predictions.

Fig. (12). High resolution *in silico* vs *in vitro* promiscuity prediction matrix. The *in vitro* vs *in silico* matrix is color coded using dark blue when there is evidence of high biological activity (IC50/nM ≤ 1), blue when there is evidence of medium biological activity (1 < IC50/nM ≤ 10), and light blue when there is evidence of low biological activity (IC50/nM > 10). The *in silico* predicted targets (Tanimoto similarity score ≥ 0.9) are highlighted as light blue bands. Analyzing the matches it can be observed that, with the available biological data, *in vitro* evidence of promiscuous targets often agrees with *in silico* predicted promiscuous targets. However, there is no experimental data for some targets that are predicted to be promiscuous, and there are some targets (mainly proteins with binding pockets on the surface which MSSH does not represent accurately) which are not predicted to be promiscuous and there is experimental evidence of it.

in vitro matrix. In both matrices, the horizontal axis is labelled according to the 76 annotations in which the 677 complexes are distributed and the vertical axis shows the 3-letter ligand code used in the promiscuity predictions. It can be seen that the available experimental data from BindingDB supports the notion that many existing targets are promiscuous (*e.g.* thrombin, coagulation factor Xa, dipeptidil peptidase IV, aldehyde reductase, dihydrofolate reductase, steryl sulfatase, cyclooxigenase 2, RNA directed DNA polymerase, hiv retropepsin). It also shows more general annotations which are expected to involve multiple targets (*e.g.* transferring alkyl groups, peptidase, metalloendo-peptidase, nitric oxide synthase or unspecific monooxy-genase).

The experimental evidence of promiscuous targets agrees well with several MDDR activity classes existent for their ligands. Supplementary Table **3** shows the promiscuous targets identified by the experimental data and the various MDDR activity classes related to their ligands.

Fig. (**12**) shows the comparison between the *in silico* and *in vitro* promiscuity predictions. The *in silico* predicted targets (Tanimoto score ≥ 0.9) are highlighted as light blue bands over the *in vitro* experimental data. Comparing the matches with the available experimental data, *in vitro* evidence of promiscuous targets often agrees with *in silico* predicted promiscuous targets (*e.g.* GABA-A alpha subunit, androgen, estrogen, acetylcholinesterase, RNA directed DNA polymerase). However, there are cases where experimental data is not available for some targets that are predicted to be promiscuous (*e.g.* procollagen proline dioxygenase, cholestenone 5α reductase, vitamin d3-like receptors, ribonucleoside diphosphate reductase), and other cases where targets are predicted not to be promiscuous but where there is experimental evidence of promiscuity (*e.g.* adenosine deaminase, metalloendopeptidase, hiv 1 retropepsin, coagulation factor Xa, thrombin, dipeptidil peptidase IV, aldehyde reductase, peptidase). This later group mainly consists of proteins with surface pockets which are not represented well in MSSH.

DISCUSSION

Previous computational studies to predict pharmaco-logical profiles have used the similarity of chemical ligand structures, protein sequences, pharmacophoric binding pockets, or phenotypic side-effects to infer whether two drugs share a target. Our novel approach relates targets by SH shape similarity in both the ligand and binding pocket spaces. This allows promiscuous ligands and targets to be predicted using an explicit shape-based representation. Since 3D shape complementarity is essential for molecular recognition, it is perhaps not surprising that our 3D shape-based approach can give very good promiscuity predictions. We have shown some specific examples in which our method can detect a promiscuous target, such as androgen, GABA-A alpha subunit, hydroxymethylglutaryl CoA reductase, and thrombin. The ligands found to be promiscuous are often small and hydrophobic, as observed previously [25]. Moreover, the binding pockets for our predicted promiscuous targets are consistent with the general requirements for promiscuity (*i.e.* large hydrophobic binding sites, evidence of alternative binding modes for the same ligand at the same site, sensitivity of the exact binding mode

to small differences in residues surrounding the site, or existence of a flexible lid to enable the binding site to accommodate a broad range of ligand sizes with good potency [52]).

As shown in previous studies [23, 51], there are cases where targets highly related by sequence are unrelated by their ligands, and cases where receptors unrelated by sequence are highly related by their ligands. Our approach has the advantage to combine information from both the ligand and binding pocket spaces. By mapping small-molecule shape space to protein binding pocket shape space, we are able to identify groups of receptors that can be unrelated by sequence and structure but which have ligands with common shapes. In this way, previously unknown cross-interactions can be detected. We have shown here that shape clustering can help to identify off-target relationships. Cross-shape matching can be used as a first approach to identify promiscuous ligands and targets, and this should save time and costs compared to using standard functional assays. This should also facilitate the search for novel targets of marketed drugs.

The work presented here has focused on a retrospective study of known MDDR drugs. We are currently working to determine a better similarity threshold and to calculate a more rigorous interaction probability.

CONCLUSION

We have presented a 3D shape-based approach for predicting drug promiscuity by correlating both ligand and binding pocket SH shapes. The method has been validated using a subset of the MDDR for which experimental information is available and has been demonstrated to be effective in identifying related targets which are known to have related MDDR activity classes.

When assessing similarity between two targets, the advantage of examining ligand-pocket shape similarity compared to protein sequence is the ability to identify targets which may have different folds but which have unexpectedly similar binding sites. Normally, polypharmacology prediction methods operate in ligand space or protein space. Comparing ligands with consensus pocket shapes leads to interesting promiscuity predictions which are often consistent with known crystallographic examples of promiscuous ligands and protein targets. Our results show that the performance in ligand space is comparable to that in binding pocket space, which provides supporting evidence for the pocket-based predictions. Moreover, the comparison of our *in silico* promiscuity predictions with the available *in vitro* results from Binding DB shows a similar agreement.

Overall, we have presented a new protocol to detect promiscuous ligands and targets and we have validated it using experimental information. Our results indicate that promiscuous ligands and targets are more common than previously assumed. Detecting and quantifying the similarities between target families will help the identification and exploitation of possible promiscuous targets.

CONFLICT OF INTEREST

Non declared.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Xie, L.; Xie, L.; Bourne, P.E. Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.,* **2011**, *21*, 1-11.

[2] Wermuth, C.G. Selective optimization of side activities: another way for drug discovery. *J. Med. Chem.,* **2004**, *47*, 1303-1314.

[3] Keiser, M.J; Irwin, J.J.; Shoichet, B.K. The Chemical Basis of Pharmacology. *Biochemistry,* **2010**, *49*, 10267-10276.

[4] Azzaoui, K. *et al*. Modeling promiscuity based on *in vitro* safety pharmacology profiling data. *Chem. Med. Chem.,* **2007**, *2*, 874-880.

[5] Merlot, C. In silico methods for early toxicity assessment. *Curr. Opin. Drug Discov. Devel.,* **2008**, *11*, 80-85.

[6] Campillos, M.; Kuhn, M.; Gavin, A.C.; Jensen, L.J; Bork, P. Drug target identification using side-effect similarity. *Science,* **2008**, *321*, 263-266.

[7] Fedorov, O. *et al*. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. USA,* **2007**, *104*, 20523-20528.

[8] Trubetskoy, O.V. *et al*. High throughput screening assay for UDP-glucuronosyltransferase 1A1 glucuronidation profiling. *Assay Drug Dev. Technol.,* **2007**, *5*, 343-354.

[9] Nobeli, I.; Favia, A.D.; Thornton, J.M. Protein promiscuity and its applications for biotechnology. *Nat. Biotechnol.,* **2009**, *27*, 157-167.

[10] Chong, C.R.; Sullivan, D.J. New uses for old drugs. *Nature,* **2007**, *448*, 645-646.

[11] Nigsch, F.; Bender, A.; Jenkins, J.L.; Mitchell, J.B.O. Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J. Chem. Inf. Model.,* **2008**, *48*, 2313-2325.

[12] Niijima, S.; Yabuuchi, H.; Okuno, Y. Cross-Target View to Feature Selection: Identification of Molecular Interaction Features in Ligand−Target Space. *J. Chem. Inf. Model.,* **2011**, *51*, 15-24.

[13] Paolini, G.V.; Shapland, R.H.B.; van Hoorn, W.P.; Mason, J.S.; Hopkins, A.L. Global mapping of pharmacological space. *Nat. Biotechnol.,* **2006**, *24*, 805-815.

[14] Weskamp, N.; Hüllermeier, E; Klebe, G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins,* **2009**, *76*, 317-330.

[15] Keiser, M. J. *et al*. Predicting new molecular targets for known drugs. *Nature,* **2009**, *462*, 175-181.

[16] Milletti, F.; Vulpetti, A. Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.,* **2010**, *50*, 1418-143.

[17] Altschul, S.A.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.,* **1990**, *215*, 403-410.

[18] Altschul, S.A.; Madden, T.L.; Scha ffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **1997**, *25*, 3389-3402.

[19] Pearson, W.R. Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol.,* **1990**, *183*, 63-98.

[20] Arnold, R.; Rattei, T.; Tischler, P.; Truong, M.D.; Stumpflen, V.; Mewes, W. Simap-the similarity matrix of proteins. *Bioinformatics*, **2005**, *21*, ii42-ii46.

[21] Johnson, M.A.; Maggiora, G.M. *Concepts and applications of molecular similarity*; Wiley & Sons: New York, **1990**.

[22] Keiser, M.J. *et al*. Relating protein pharmacology by ligand chemistry. *Nature Biotechnol.,* **2007**, *25*, 197-206.

[23] Hert, J.; Keiser, M.J.; Irwin, J.J.; Oprea, T.I.; Shoichet, B.K. Quantifying the relationships among drug classes. *J. Chem. Inf. Model.,* **2008**, *48*, 755-765.

[24] Vidal, D.; Mestres, J. In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.,* **2010**, *29*, 543-551.

[25] Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R.V. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. BioSyst.,* **2009**, *5*, 1051-1057.

[26] 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.

[27] 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.

[28] Schmitt, S.; Kuhn, D.; Klebe, G.A. New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homol- ogy. *J. Mol. Biol.,* **2002**, *323*, 387-406.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Belongie, S.; Malik, J.; Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2002**, *24*, 509-522.

[34] Richmond, N.J.; Willett, P.; Clark, R.D. Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graphics Model.,* **2004**, *23*, 199-209.

[35] Lin, J.; Clark, T. An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.,* **2005**, *45*, 1010-1016.

[36] Ritchie, D.W.; Kemp, G.J.L. Protein Docking Using Spherical Polar Fourier Correlations. *Proteins: Struct. Func. Genet.,* **2000**, *39*, 178-194.

[37] Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J-J.; Lecchini, L.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 947-955.

[38] MDL Drug Data Report, 2010.2 (MDL Informations Systems Inc., San Leandro, CA, **2010**).

[39] Berman, H.M.;Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissing, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.,* **2000**, *28*, 235-242.

[40] Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R.N.; Gilson, M.K. Binding DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.,* **2007**, *35*, D198-D201.

[41] CEPOS InSilico Ltd.: Erlangen, Germany, 2009. http://www.ceposinsilico.de/ (accessed: June 24, **2009**).

[42] Beautrait, A.; Leroux, V.; Chavent, M.; Ghemtio, L.; Devignes, M.-D.; Smaïl-Tabbone, M.; Cai, W.; Shao, W.; Moreau, G.; Bladon, P.; Yao, Maigret, B. J. Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment. *J. Mol. Model.,* **2008**, *14*, 135-148.

[43] Mavridis, L.; Hudson, B.D.; Ritchie, D.W. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *J. Chem. Inf. Model.,* **2007**, *47*, 1787-1796.

[44] Pérez-Nueno, V.I.; Ritchie, D.W.; Borrell, J.I.; Teixidó, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape based virtual screening approach: Further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.,* **2008** , *48*, 2146-2165.

[45] Pérez-Nueno, V.I.; Ritchie, D.W.; Rabal, O.; Pascual, R.; Borrell, J.I.; Teixidó, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape-matching and ligand-receptor docking. *J. Chem. Inf. Model.,* **2008**, *48*, 509-533.

[46] Venkatraman, V.; Pérez-Nueno, V.I.; Mavridis, L.; Ritchie, D. W. A comprehensive comparison of ligand-based virtual screening reveals limitations of current 3D methods tools against the DUD dataset. *J. Chem. Inf. Model.,* **2010**, *50*, 2079-2093.

[47] Pérez-Nueno, V.I.; Ritchie, D.W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening. *J. Chem. Inf. Model.*, **2011**, *51*, 1233-1248.

[48] Pérez-Nueno, V.I.; Venkatraman, V.; Mavridis, L.; Clark, T; Ritchie D.W. Using spherical harmonic surface property representations for ligand-based virtual screening. *Mol. Inf.,* **2010**, *30*, 151-159.

[49] International Union of Biochemistry and Molecular Biology, Nomenclature Committee & Webb, E.C. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union Of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*; Academic Press: San Diego, **1992**.

[50] Ritchie, D.W.; Kemp, G.J.L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comput. Chem.*, **1999**, *20*, 383-395.

[51] Adams, J.C.; Keiser, M.J.; Basuino, L.; Chambers, H.F.; Lee, D.S.; Wiest, O.G.; Babbitt, P.C. A mapping of drug space from the viewpoint of small molecule metabolism. *PLoS Comput. Biol.,* **2009**, *5*, e1000474, 1-12.

[52] Hopkins, A.; Mason, J.S.; Overington, J.P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.*, **2006**, *16*, 127-136.