

Chemoisosterism in the Proteome

Xavier Jalencas and Jordi Mestres*

Chemogenomics Laboratory, Research Programme on Biomedical Informatics (GRIB), IMIM Hospital del Mar Research Institute and University Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

Supporting Information

ABSTRACT: The concept of chemoisosterism of protein environments is introduced as the complementary property to bioisosterism of chemical fragments. In the same way that two chemical fragments are considered bioisosteric if they can bind to the same protein environment, two protein environments will be considered chemoisosteric if they can interact with the same chemical fragment. The basis for the identification of chemoisosteric relationships among protein environments was the increasing amount of crystal structures available currently for protein–ligand complexes. It is shown that one can recover the right location and orientation of chemical fragments constituting the native ligand in a nuclear receptor structure by using only chemoisosteric environments present in enzyme structures. Examples of the potential applicability of chemoisosterism in fragment-based drug discovery are provided.



INTRODUCTION

The amount of protein structure information available in the public domain continues to grow exponentially, and today there are over 72,000 X-ray structure entries in the Protein Data Bank (PDB).¹ Not only the number of structures is increasing but their functional coverage is being expanded as well. There is still a strong bias for therapeutic targets,² but it is slowly being corrected thanks to recent structural genomics initiatives.³ Consequently, even though the main body of information on protein structures is still devoted to enzymes,² members of protein families traditionally difficult to crystallize, such as G protein-coupled receptors (GPCRs) and other membrane proteins, have seen their first structures crystallized in recent years.⁴

Many of the protein crystal structures available contain a ligand forming a complex inside a binding cavity.⁵ This information provides the ultimate structural confirmation of the degree of compatibility between the chemical fragments present in bioactive molecules and the protein environments shaping the binding cavity.⁶ In this respect, the property by which different chemical fragments can be accommodated in the same protein environment without penalizing strongly the bioactivity of a given small molecule is usually referred to as bioisosterism.⁷ This is schematically illustrated in Figure 1a. Within a given protein environment, some bioisosteric chemical fragments may show relatively high structural similarities (e.g., imidazole and tetrazole), whereas the structures of others can be more distantly related (e.g., imidazole and carboxylate).⁸ Applying bioisosteric transformations to parts of the structure of a bioactive compound is a widely used strategy in medicinal chemistry to optimize both the pharmacokinetic and pharmacodynamic profiles of molecules or to explore the chemical space around protected chemical series.^{9–11}

Bioisosterism looks at the interaction between chemical fragments and protein environments from the perspective of the latter. If the reference framework is shifted to chemical

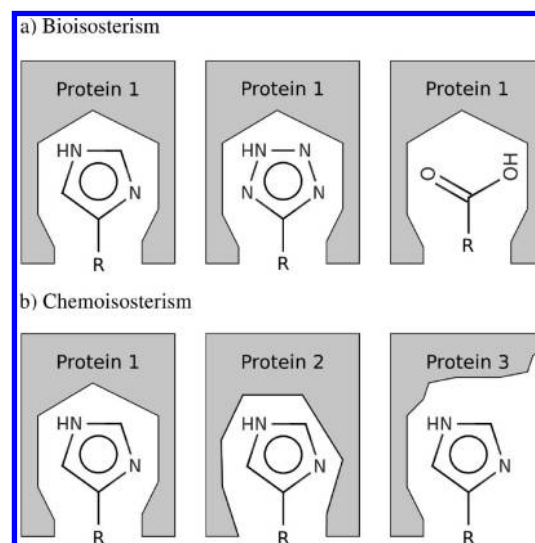


Figure 1. Schematic definition of (a) bioisosterism and (b) chemoisosterism.

fragments, the corresponding complementary concept emerges. Accordingly, the property by which different protein environments can tolerate the interaction with the same chemical fragment will be referred to as chemoisosterism. Figure 1b shows three chemoisosteric protein environments compatible with the interaction with an imidazole group. As mentioned above for bioisosterism, some chemoisosteric protein environments may share similar cavity shapes and features (e.g., proteins 1 and 2), whereas others can be genuinely different (e.g., proteins 1 and 3).

Received: June 26, 2012

Published: January 11, 2013

Within this context, gaining a deeper understanding of the degree of chemoisosterism among protein environments can shed some light into the inherent cross-pharmacology between protein targets,^{12,13} as well as into the selectivity of bioactive small molecules and its further implications to toxicity and adverse reactions.^{14–17} Accordingly, this work aims at exploring the chemoisosteric relationships between protein environments that can be directly inferred from the currently available structures of protein–ligand complexes. An application example to identifying chemical fragments present in molecules with known affinity for a nuclear receptor using protein environments from enzymes only is presented and its broader impact for fragment-based drug discovery discussed.

METHODS AND MATERIALS

Extraction of Chemical Fragments and Protein Environments. Crystal structures of protein–ligand complexes were obtained from the PDBbind database,¹⁸ a public resource that provides a comprehensive collection of experimentally measured binding affinity data for biomolecular complexes deposited in the PDB.¹ In its 2009 version, it contains 4277 protein–ligand entries with known binding data, almost 80% of which being enzyme structures. Therefore, it ought to be stressed that the chemoisosteric relationships extracted from the data present in PDBbind will be subject to the inherent bias of the currently available protein structure information, with entries from GPCRs still being largely under-represented.

As a first step, all ligands were fragmented. Cleavable bonds within each molecule were recognized using some predefined rules.¹⁹ The list of fragments extracted from each molecule is ultimately generated by applying all possible combinations of bond cleavage. This process resulted in a total of 112,517 chemical fragments containing less than 20 heavy atoms. From the protein side, the SMART software was used to define a triangulated surface of the binding cavity.²⁰ For each chemical fragment in a ligand, its interacting protein environment was defined by all surface points that lie within a 1 Å distance from the van der Waals radii of any of the fragment atoms. In the current implementation, water molecules present in the binding site are not considered. Subsequently, each protein environment surface vertex was labeled according to the pharmacophoric features corresponding to the atom types of the underlying atoms, namely, hydrophobic, aromatic, hydrogen bond acceptor, and hydrogen bond donor. Those labeled surface vertices were grouped in surface patches (being each patch defined here as set of connected vertices assigned to the same pharmacophoric property) according to the following set of rules. For hydrophobic surface patches, all connected hydrophobic surface vertices assigned to the same heavy atom were grouped in the same surface patch. Likewise, all aromatic surface vertices assigned to the same aromatic group of atoms generated an aromatic surface patch, but in this case, only the surface vertices for which the angle between the normal vector of the aromatic group and the vector from the vertex to its corresponding atom is below 60° were considered. This filtering ensures that only the surface vertices at both sides of an aromatic group generate aromatic patches. A similar filtering was applied to catch the directionality of hydrogen bonding. Only surface vertices in the direction of a theoretical hydrogen bond interaction (allowing a deviation of 30°) were considered when forming hydrogen bond acceptor and donor patches.

Following a similar strategy described in previous works,^{21,22} regions of surface vertices annotated to the same pharmacophoric feature

were finally condensed into a sole surface feature point. Accordingly, for each surface patch, a surface feature point of the same pharmacophoric type was created at the position of the surface vertex closest to the center of mass of all vertices in the surface patch. The set of surface feature points and their respective distances constitute a graph representation of the 112,517 protein environments that interact with the same number of chemical fragments. At this stage, in order to remove all the small and almost featureless environments, only those having five or more surface feature points were retained, resulting in a total of 87,711 protein environments with their respective chemical fragments. Unlike other approaches that collect and compare full ligand binding cavities or pockets,^{23–25} the protein environments defined here focus on the interaction with small chemical fragments and can have discontinuous surface regions, but it ought to be mentioned also that the present implementation of the approach depends on the exact patching of the binding site surface, which can produce slightly different patterns of feature surface points and thus result in essentially different protein environments. Therefore, the use of an ensemble of different structures and ligands with different fragmentation patterns is recommended.

Comparison of Protein Environments. Pairwise comparisons of protein environments are performed by counting the number of shared surface feature points (graph nodes) in the largest common subgraph. This is done by building a product graph, followed by a clique detection algorithm.^{21,26} Each node in the product graph corresponds to a feature-matching pair of points from each of the protein environments under comparison. Product graph nodes are connected by an edge if the distances between the original surface feature points in both protein environments are similar within a given threshold (<1 Å). The use of this threshold incorporates a certain degree of fuzziness when comparing protein environments that helps alleviate the effect on the protein surface of slight backbone movements and side chain conformational changes upon binding of different ligands. Once the product graph is built, the Cliquer software is used to find the largest maximal cliques (maximal fully connected subgraphs) that correspond to the largest common subgraphs between the original graphs.²⁶ This process, if applied to comparing two protein binding sites having on average 93 feature points, can be extremely demanding computationally. In contrast, the computational cost of comparing two protein environments that have on average 15 feature points is more assumable. The outcome is an ensemble of possible pairings between surface feature points that can be used to superimpose the two protein environments by minimizing their root-mean-square deviation. At this stage, a filter is applied to the 100 largest maximal cliques to ensure that their surfaces after superimposition expose the site interacting with the respective chemical fragments in the same direction.²¹ The largest remaining clique passing the surface directionality filter is selected as the best solution. The final similarity score between two protein environments I and J is computed using the cosine metric $F_{IJ}/(F_I^{1/2} \cdot F_J^{1/2})$, where F_{IJ} is the number of matched feature points in the best clique solution and F_I and F_J are the number of feature points in each protein environment.²⁷

The process described above is also applicable to apo protein structures. In this case, the complete binding site identified in the apo structure is compared to all protein environments linked to chemical fragments extracted from the PDBbind using the same similarity score defined above applied locally on the surroundings of the protein environment. In addition, it is also possible to use

protein environments containing a docked ligand structure or a dummy composite structure derived from the individual structures of several co-crystallized ligands. In the latter case, because no clear chemical fragments may be identified, the dummy structure can be used to define a composite binding site, and the process will then proceed as in the case of the apo structure.

Data Management and Storage. A total number of 87,711 protein environments (described by their surface feature points) associated with their corresponding chemical fragments (described by their chemical structure) collected from the PDBbind¹⁸ database needed to be stored in a manageable and flexible manner. To this aim, a chemical graph identifier (CGI) was generated for each of the chemical fragments, leading to 16,533 unique chemical fragments.²⁸ Of those, 14,332 are bound to enzyme environments, 8893 of which contain at least one ring, one heteroatom, and at most one chiral center, features all considered of particular relevance for bioactivity.²⁹ Among the 16,533 unique chemical fragments, there are 1072 fragments that are composed of a single ring system, of which 935 are bound to enzyme protein environments, and from those, 731 contain at least one ring, one heteroatom, and at most one chiral center. All these sets of chemical fragments were stored as sd files containing the CGI, the type, and three-dimensional coordinates of all atoms in the fragment, and additional fields with information on their interacting protein environments and corresponding protein structures.

With respect to protein environments, the same CGI algorithm applied to the chemical fragments was adapted to generate a protein environment identifier (PEI) for each of the 87,711 protein environments mapped originally to chemical fragments.²⁸ Grouping all protein environments by their PEI leads to a total of 73,931 unique protein environments, of which 71,299 correspond to enzyme protein environments and 7767 are linked to chemical fragments composed of a single ring system. However, as mentioned above, slight backbone movements and side chain conformational changes may affect the position of the surface feature points and thus lead to a different PEI for essentially the same protein environment. To circumvent this effect, an all-against-all pairwise comparison of the 7767 unique protein environments was performed. A hierarchical agglomerative clustering on the resulting similarity matrix lead to a final number of 3177 protein environment clusters of particular biological relevance that could be considered essentially unique.³⁰ An illustrative example of one of those clusters containing 13 protein environments from various serine proteases is presented in Figure 2. As observed, even though the number of surface feature points varies among the 13 environments collected in this cluster, they all share on average a set of 11 surface points that makes them significantly more similar among them than to any of the other protein environments considered. Finally, all protein environments were stored as a SD file containing the PEI, a list of all other environments in their cluster, the pharmacophore type and three-dimensional coordinates of all surface feature points in the environment, and additional fields with information on the interacting chemical fragments (CGIs) and corresponding protein structures (PDB codes).

RESULTS AND DISCUSSION

Construction of a Fragment-Environment Interaction Network. The use of network-based approaches to visualize and analyze complex interaction data has become increasingly

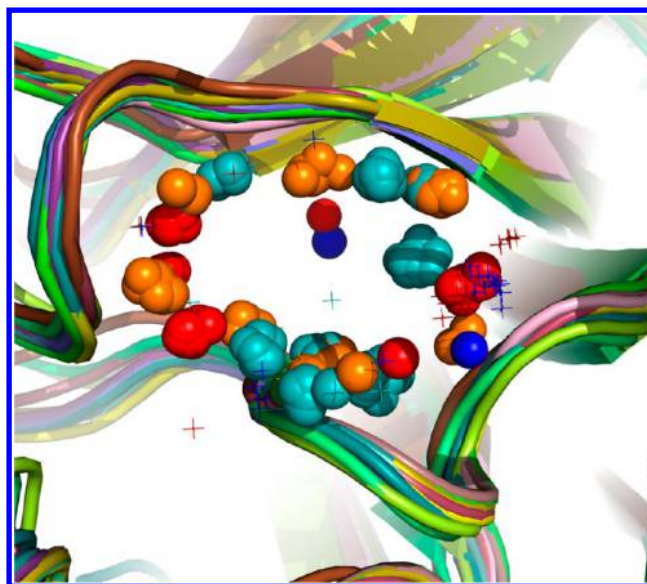


Figure 2. Cluster containing 13 protein environments found in crystal structures of serine proteases. The environments were extracted from the following list of PDB codes: 1d6w, 1dwd, 1o2g, 1owd, 1owe, 1owh, 1sqa, 1tnk, 1uto, 2vio, 2viq, 2viv, and 2vnt. Color coding of surface feature points: green (hydrophobic), orange (aromatic), red (acceptor), and blue (donor).

popular in recent years.³¹ In particular, bipartite networks capturing fragment-environment interactions are the type of emerging tools that can become extremely useful in medicinal chemistry as they inherently contain information on both bioisosteric and chemoisosteric relationships among chemical fragments and protein environments, respectively. On one side, protein environments connected to a large number of chemical fragments identify interaction sites that can tolerate multiple chemical modifications and thus can be considered *bioisosteric hotspots* in protein structures, likely associated with the ligandability of proteins and its local conservation.^{29,32} On the other side, chemical fragments connected to a large number of protein environments identify functional groups that can be easily accommodated into multiple interaction sites from a variety of protein structures and thus can be considered *chemoisosteric hotspots* in chemical structures, likely associated with promoting polytarget pharmacology in ligands.^{14,15,33}

The number of rings present in a small molecule is an important descriptor for drug-like molecules.³⁴ Accordingly, focus will be given to the 1072 chemical fragments composed of a single ring system and the corresponding 3177 clusters of interacting protein environments. Figure 3 shows the complex network of interactions between those chemical fragments and protein environments. In this network, a chemical fragment node is linked to a protein environment node if evidence of its interaction exists in crystal structures of protein–ligand complexes.¹⁸ Far from having a well-organized modular structure, the largest connected component of this bipartite network shows a highly complex topology, typical of cases where nodes of one class have multiple interactions with nodes of the other class.³⁵ It contains 3424 nodes involving 677 chemical fragments (in black) and 2747 clusters of protein environments (in color), of which 2324 correspond to protein environments found in enzyme structures. The majority of these enzyme clusters (2091) are composed of a consistent set of environments that belong to the same enzyme subclass and

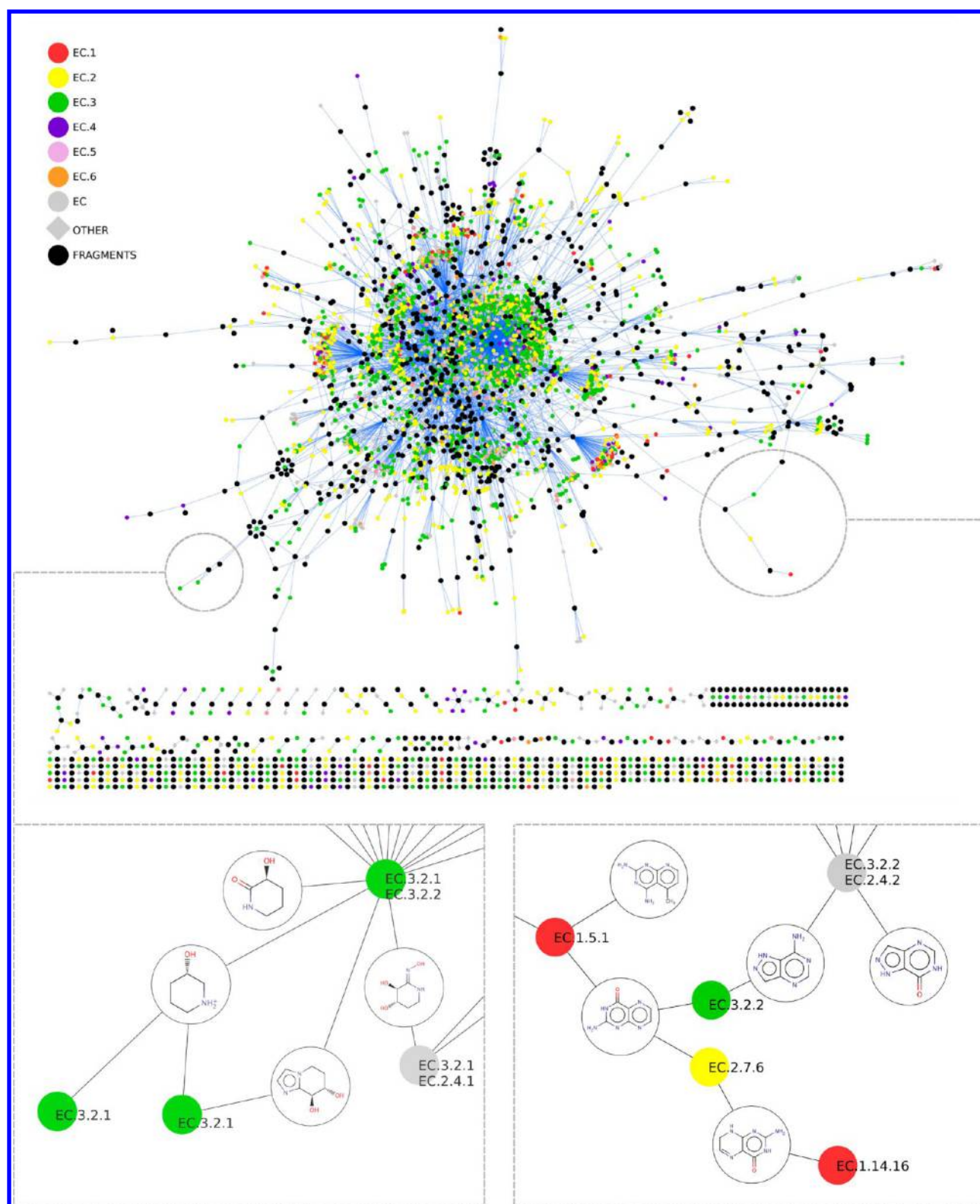


Figure 3. Bipartite interaction network between chemical fragments (black dots) and protein environments (colored dots) extracted from PDBbind. The bioisosteric and chemoisosteric relationships between chemical fragments and protein environments, respectively, are shown in detail for two sample regions of the network.

are thus colored according to that subclass. However, some enzyme clusters (233) agglomerate protein environments from various enzyme subclasses, and a generic EC label is then assigned to them (gray circles). There are also 423 protein environments that come from protein entries other than

enzymes (OTHER), of which a portion (66) are structures of nuclear receptors.

Two sample regions of the network are magnified in Figure 3 to illustrate some of the bioisosteric and chemoisosteric relationships that emerge directly from protein structure

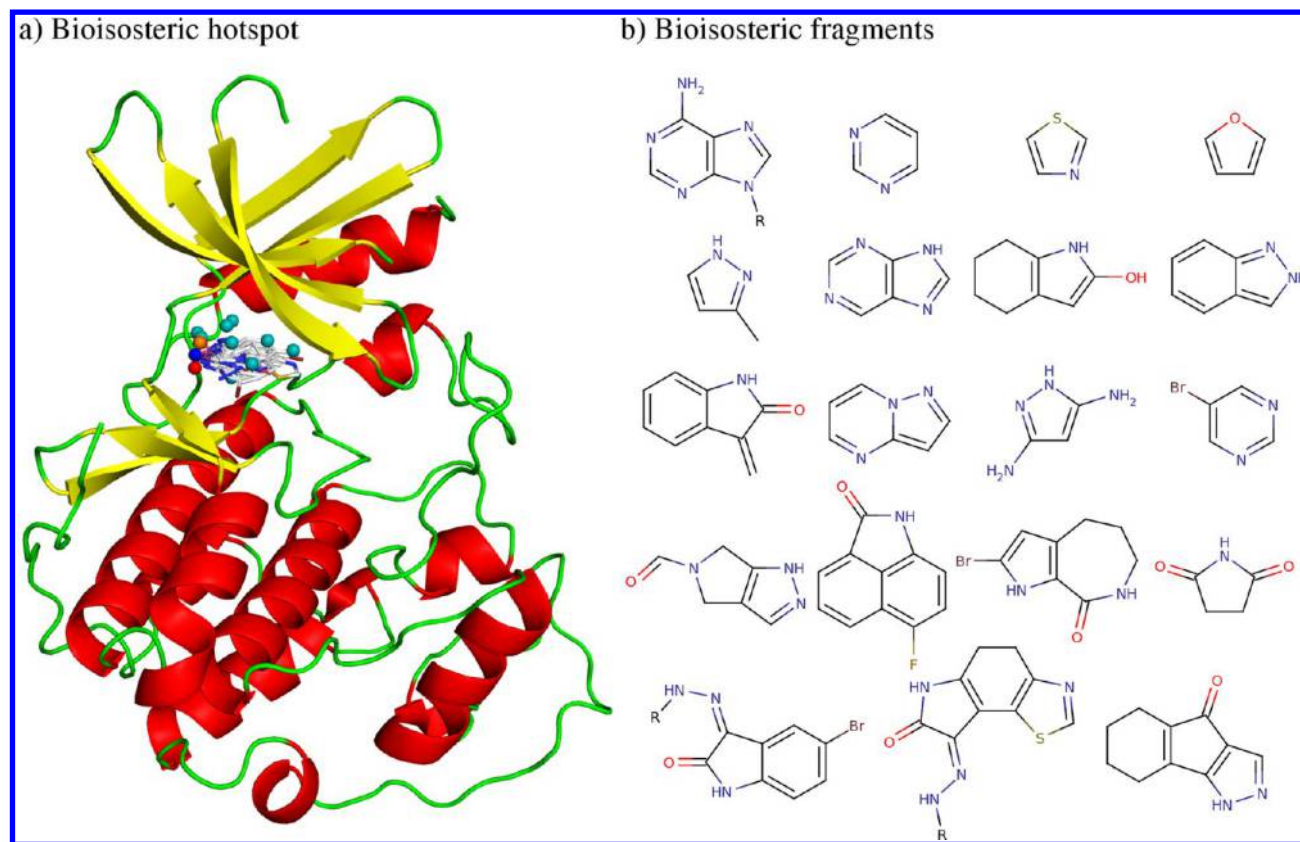


Figure 4. (a) Bioisosteric hotspot identified in 28 structures of 4 different kinases displayed in a structure of cdk2 as template (PDB code 1b38) and (b) the 19 single-ring bioisosteric fragments that interact in this protein environment.

information. On the left, a region rich on bioisosteric relationships between a set of 3-hydroxypiperidin fragments and clusters of protein environments from glycosidases (EC 3.2.1) is presented, whereas on the right focus is given to a region rich on chemoisosteric relationships between clusters of protein environments from oxidoreductases acting on CH–NH group donors and with NAD⁺ or NADP⁺ as acceptor (EC 1.5.1), diphosphotransferases (EC 2.7.6), and glycosidases hydrolyzing N-glycosyl compounds (EC 3.2.2) and a pterin fragment.

Identification of Bioisosteric and Chemoisosteric Hotspots. A more detailed analysis of the fragment–environment interaction network presented in Figure 3 reveals the existence of both bioisosteric and chemoisosteric hotspots, that is, protein environments that can accommodate a large number of chemical fragments and chemical fragments that can interact with a large number of protein environments, respectively. The protein environment node with the largest number of links to chemical fragments is a cluster composed of 28 protein environments from 4 kinases (24 from cdk2, 2 from Aurora-A, 1 from chk1, and 1 from gsk3) connected to 19 unique fragments composed of a single ring system. Figure 4a shows that this bioisosteric hotspot is located at the heart of the ATP site of those kinases, a recognized essential region of interaction in this protein class.³⁶ The majority of fragments linked to this environment (12 out of 19) share the same hydrogen bond interaction pattern with residues Glu81 and Leu83 of cdk2 (PDB code 1b38). The structures of the 19 chemical fragments are also provided in Figure 4b, ordered from left to right and top to bottom according to the number of links to protein environments. Among them, the adenylyl fragment is the one linked to the largest number of chemoisosteric environments (102). It is

present in 4.8% of the ligands in PDBbind, but it is also comparatively present in 2.8% of drugs³⁷ and 0.5% of all bioactive compounds in public sources of pharmacological data.^{38,39}

Among the top bioisosteric hotspots identified, there are two protein environment nodes from serine proteases. A cluster composed of 41 protein environments from thrombin structures occupies rank number 2. It is connected to 17 unique chemical fragments composed of a single ring system that all share a hydrogen-bonding interaction with Gly216, a key residue that helps anchor the structure of ligands in the region that faces the oxyanion hole (Figure 1, Supporting Information). In addition, rank number 3 corresponds to 117 protein environments from 7 serine proteases (68 from trypsin, 18 from factor Xa, 16 from U-plasminogen activator, 9 from thrombin, 3 from factor VIIa, 2 from matriptase, and 1 from T-plasminogen activator) connected to 17 unique chemical fragments composed of a single ring system that interact in the well-conserved S1 pocket of the active site of these enzymes (Figure 2, Supporting Information).

Conversely, the top four chemoisosteric hotspots containing at least one heteroatom are phenol, 3,4-dihydroxyfuran, pyridine, and chlorophenyl that are connected to 999, 160, 148, and 118 protein environments, respectively. The presence of these chemical fragments in molecular structures should enhance the chances of having affinity for multiple targets. A phenol group, for example, is found in 7.0% of ligands in PDBbind interacting with 57 different proteins, but it is also present in 4.7% of drugs and 4.0% of small molecules present in public sources of pharmacological data having relevant bioactivity (pK_b , pK_d , pIC_{50} , or $pEC_{50} > 6$). In turn, 3,4-dihydroxyfuran, pyridine, and chlorophenyl are found in 6.6%, 5.1%, and 3.8% of ligands in

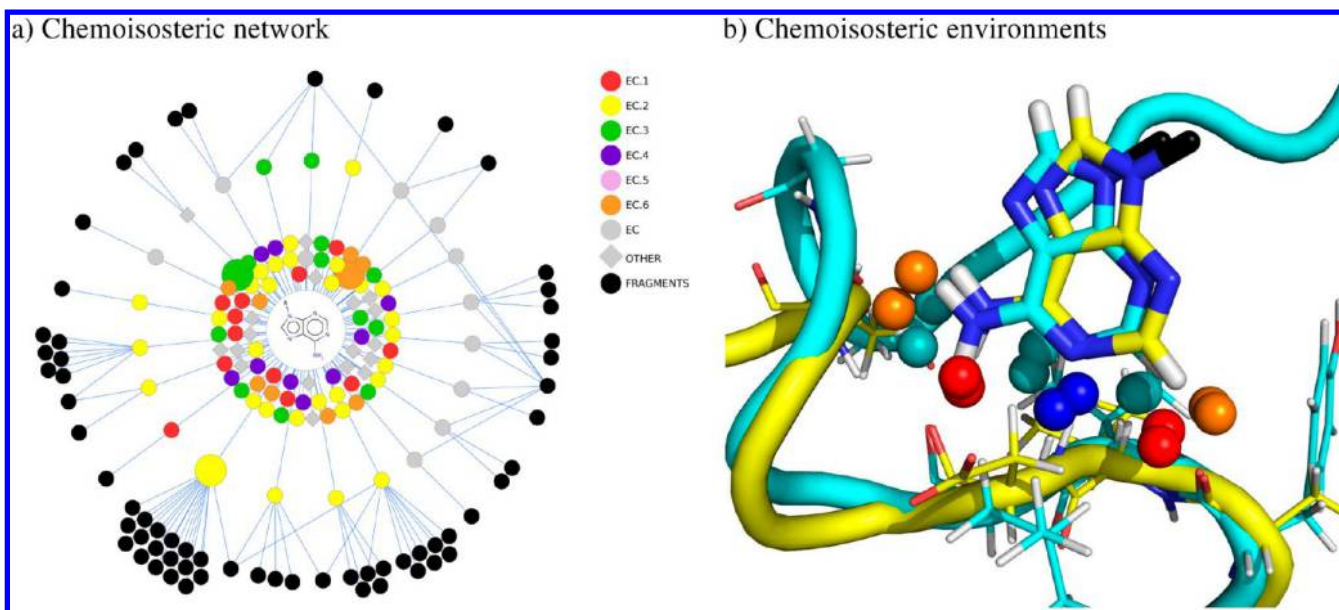


Figure 5. (a) Chemoisosteric network built around the adenyl fragment and (b) an illustrative example of a pair of chemoisosteric protein environments from structurally diverse proteins that interact with the adenyl fragment, namely, aminoacyl-tRNA synthase (PDB code 2v0c, cyan) and ATP synthase (PDB code 2e5y, yellow).

PDBbind interacting with 58, 49, and 49 different proteins, respectively.

Among the top chemoisosteric hotspots identified, the adenyl fragment occupies rank number 7. Figure 5a illustrates the chemoisosteric network built around the adenyl fragment. A total of 166 nodes are shown in this network. Apart from the central node displaying the structure of the adenyl fragment, the rest of the nodes are distributed in three layers, the first two with 102 protein environments (colored dots), 83 of them from enzymes, and the third one with 63 chemical fragments (black dots). The first layer contains the 80 interacting protein environments unique to adenyl. As already commented (Figure 1), some of the chemoisosteric environments to adenyl may be quite diverse in the number of surface feature points and their arrangement, but others can have similar surface feature points arranged similarly. As an example, Figure 5b shows the superposition of two protein environments from structures of aminoacyl-tRNA synthetase (PDB code 2v0c, EC 6.1.1.4, in cyan) and an ATP synthase (PDB code 2e5y, EC 3.6.3.14, in yellow), highlighted as large green and orange circles, respectively, in the first layer of the network in Figure 5a. In spite of being present in two proteins that belong to different enzyme subclasses and appear to be quite diverse structurally in this site, they locally share eight surface feature points that are linked to the interaction with an adenyl fragment. Accordingly, the identification of similar protein environments, using some of the clique detection approaches described previously, can be a useful approximation to infer potential chemoisosteric relationships among protein environments of unrelated proteins. The second layer contains the 22 protein environments that interact with another chemical fragment in addition to adenyl. The cluster of protein environments from four kinases depicted in Figure 4 is marked with a large yellow dot in this second layer. Finally, the third layer depicts all chemical fragments that are found to interact with protein environments that in turn interact with the adenyl fragment and are thus bioisosteric to adenyl.

It ought to be stressed here that one of the caveats of the approach is that many chemical fragments associated to protein

environments in our database may not bind in the same relative position and orientation, or even not bind at all, when considered in isolation, disengaged from the rest of the ligand structure from which it was extracted. However, the existence of structural evidence of their mutual tolerance is undeniable. For that reason, we will speak of chemical fragments being compatible with protein environments, not necessarily implying that they are particularly good binders or privileged to those protein environments.

Inferring Chemoisosterism under Similarity Constraints. Even though bioisosteric relationships between chemical fragments are not bound by their similarity (that being based on molecular properties, structures, or pharmacophoric features), it is often assumed that the identification of similar chemical fragments is a good approximation to infer bioisosterism.⁸ Along the same lines, although it has been shown that chemoisosteric relationships between protein environments are not necessarily bound by their similarity (Figure 5a), one can assume that identifying protein environments that expose a similar network of interacting features (Figure 5b) is an efficient strategy to infer chemoisosterism.⁴⁰ Accordingly, an automatic procedure was developed to generate a chemical fragment mapping of a complete protein binding site by inferring chemoisosteric relationships between the protein environments extracted from a crystal structure of a protein–ligand complex and those found in distantly related proteins. The basic premise is that protein environments that share a common bound chemical fragment will likely share others, a premise that was recently found to apply to a large extent on protein–chemical complexes.⁴¹ The whole process was applied to the binding cavity of a nuclear receptor, the peroxisome proliferation activated receptor gamma (PPAR γ), using only chemoisosteric relationships with protein environments present in enzyme structures.

To illustrate the process, a crystal structure of a nanomolar affinity ligand (GI262570) inside the binding cavity of PPAR γ was selected from the PDB (1fm9, chain D). Initially, the ligand is identified in the crystal structure and the corresponding protein surface of interaction derived (Figure 6a). Fragmentation rules

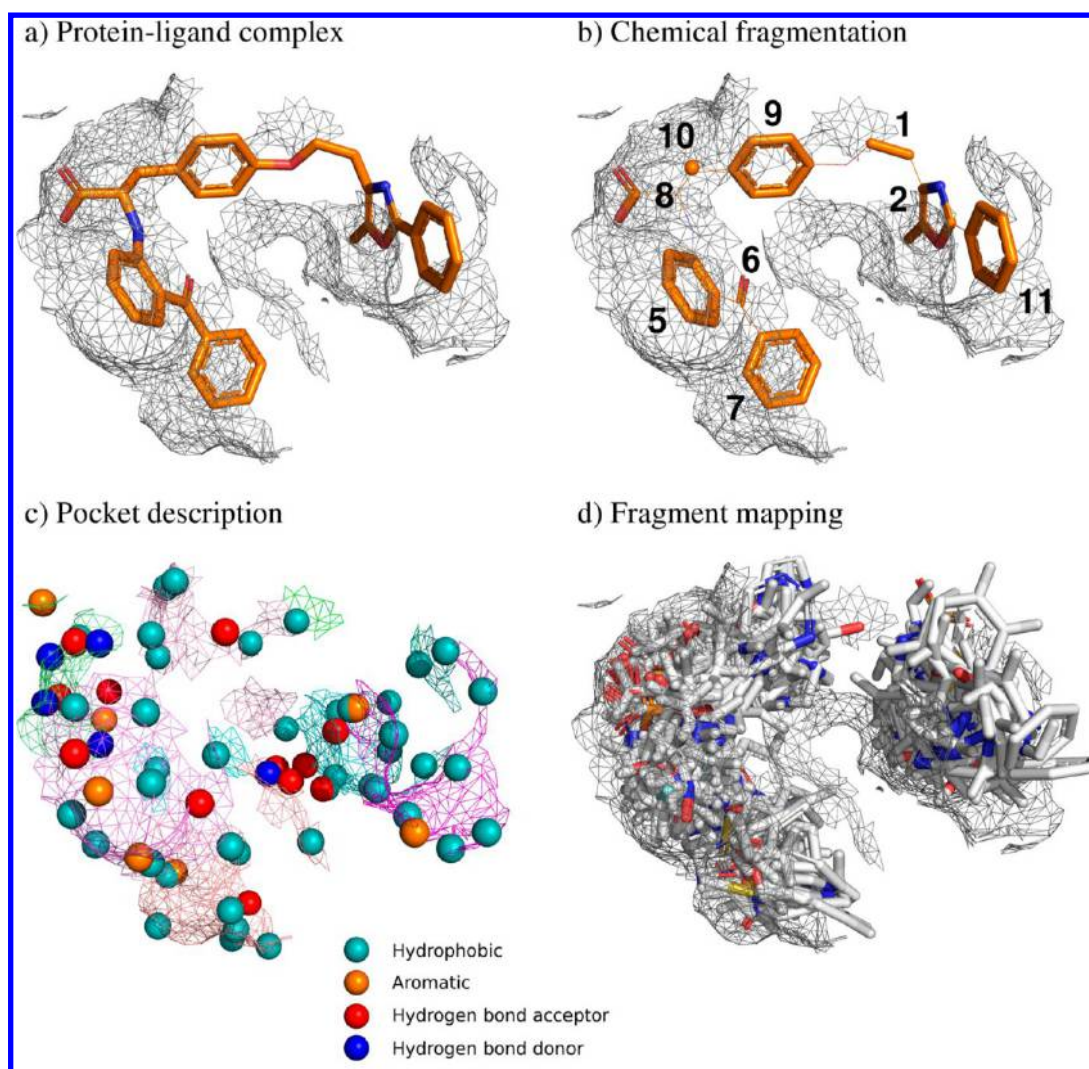


Figure 6. Process of inferring chemoisosterism from distantly related proteins: (a) starting from a structure of a protein–ligand complex, (b) chemical fragments are obtained according to ligand fragmentation rules, from which (c) the corresponding interacting protein environments are defined by surface features points are used (d) to identify similar protein environments in other structures with known interacting chemical fragments that are directly mapped by superimposition into the binding cavity of the original structure. The example shown corresponds to the interaction of GI262570 with PPAR γ (PDB code 1fm9).

are then applied to the ligand, and protein environments are assigned to each of the resulting chemical fragments (Figure 6b). In total, 12 chemical fragments were generated at this stage. Afterward, protein environments are segmented according to four pharmacophoric features, and representative surface feature points are generated (Figure 6c). Only those chemical fragments that can be mapped to protein environments described by at least five surface feature points are retained. In this case, 7 out of the 12 original chemical fragments passed that filter (fragments 2, 5, 7, 8, 9, 10, and 11 in Figure 6b). Finally, each of the corresponding seven protein environments in PPAR γ was compared against the 71,299 unique protein environments extracted from enzyme structures in PDBbind. This was done to ensure that only chemoisosteric relationships between protein environments of distantly related proteins could be inferred. When a protein environment in an enzyme structure was found similar to any of the seven protein environments in PPAR γ , the match between their surface feature points was used to superimpose them and map directly the chemical fragment interacting with the enzyme environment on to the PPAR γ environment. It is at this stage where a chemoisosteric relationship is inferred through the

assumption that if two protein environments are similar and one of them is known to interact with a given chemical fragment, chances are that it may interact also with the other protein environment. To ensure that all chemical fragments that were mapped directly, based solely on the superposition of protein environments, fit reasonably well inside the PPAR γ cavity, an interaction energy was obtained with AutoDock.⁴² All chemical fragments having positive energy values were ultimately removed. In the end, a PPAR γ cavity filled with chemical fragments from potentially chemoisosteric enzyme environments was obtained (Figure 6d). In this respect, the final outcome is reminiscent of the results obtained from a traditional multicopy simultaneous search (MCSS).⁴³ This information can now be exploited in fragment-based drug discovery, from lead generation to fragment screening.⁴⁴

Are the Right Fragments Found in the Right Place? Provided that a crystal structure of our target protein is available, the ability to identify the right chemical fragments in the correct region of the binding cavity can contribute significantly to lead generation programs, both in guiding synthetic efforts to optimize an original fragment hit and in suggesting structural

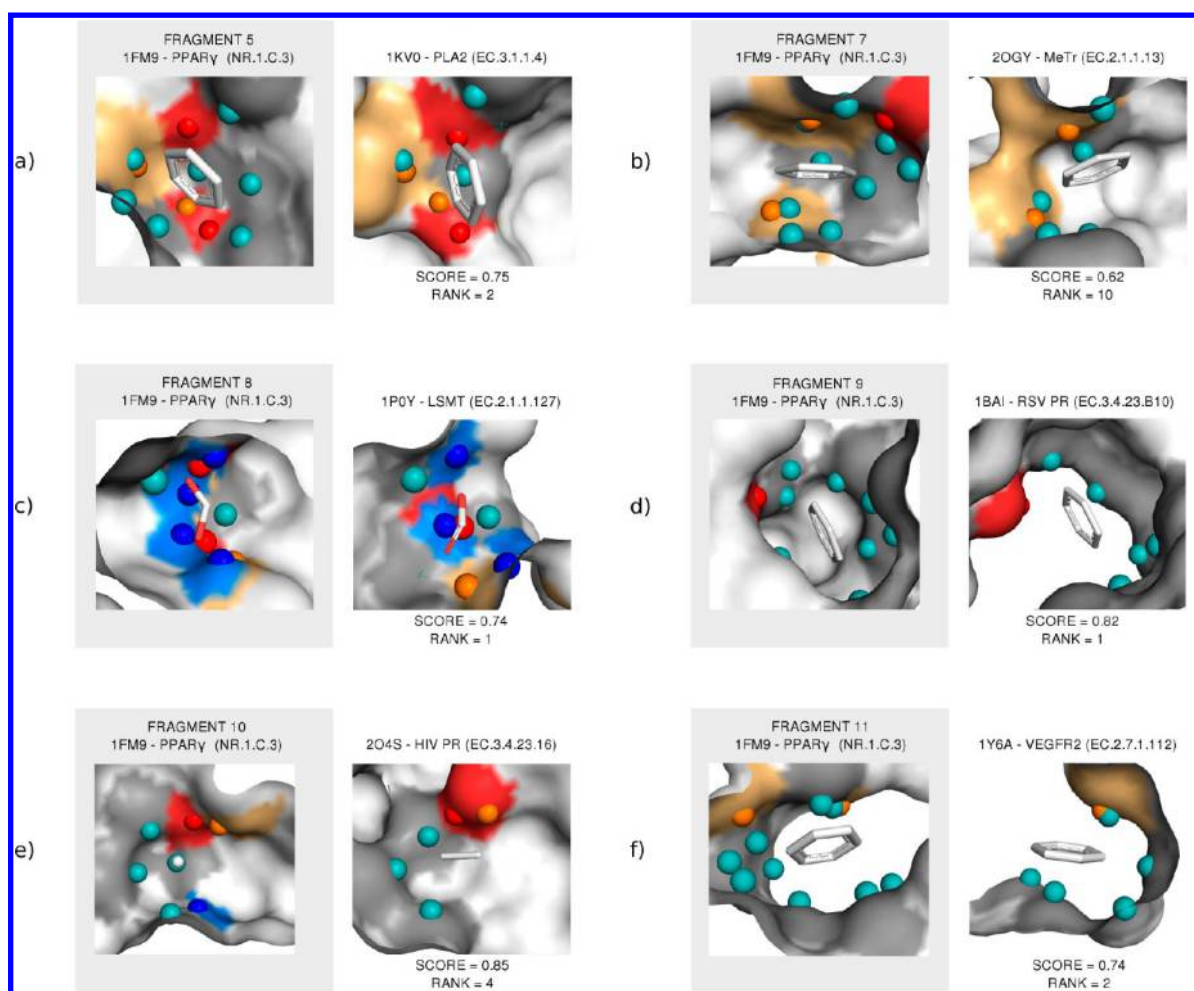



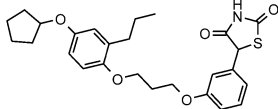

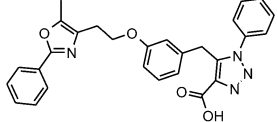
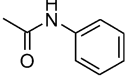
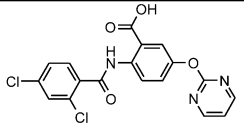

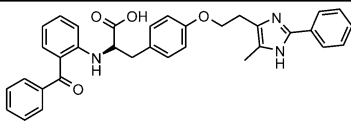
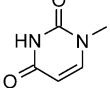

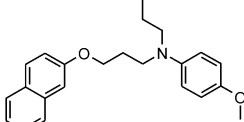
Figure 7. Set of enzyme environments chemoisosteric to PPAR γ environments. (a) PLA2: phospholipase A2. (b) MeTr: 5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase. (c) LSMT: [ribulose-bisphosphate carboxylase]-lysine N-methyltransferase. (d) RSV PR: Rous sarcoma virus protease. (e) HIV PR; HIV-1 protease. (f) VEGFR2: vascular endothelial growth factor receptor 2. Fragment numbering is consistent with Figure 6b. The score is the cosine-like similarity between the PPAR γ and the enzyme environment. The rank is the overall ranking position of the fragment shown among all fragments mapped to the PPAR γ environment.

elements to design de novo new bioactive ligands.⁴⁵ Accordingly, a key validation step for the utility of the fragment mapping obtained in the previous section (Figure 6d) is to examine whether those chemical fragments present in the native ligand of the PPAR γ structure (PDB code 1fm9) are recovered in the same region of space in about the same binding orientation. The analysis is provided in Figure 7. As observed, out of the seven chemical fragments for which feature-rich protein environments could be assigned (vide supra), the approach is able to retrieve among the top solutions the exact same chemical fragment in approximately the same location and orientation for six of them, namely, the four phenyl groups (fragments 5, 7, 9, and 11), the carboxylate group (fragment 8), and the methylene fragment (fragment 10) as depicted in Figure 6b. The only exception was the methyl oxazole (fragment 2), for which only the methyl group could actually be reproduced in the same location. Being a pure knowledge-based approach, this limitation is mainly due to the low representativity of this particular chemical fragment in the co-crystallized ligands present in PDBbind, with only four instances mapped to protein environments that could not be matched with the corresponding environment in PPAR γ .

A look at the degree of resemblance between the pairs of protein environments shown in Figure 7 provides a good visual

appreciation of the chemoisosteric relationships inferred between the 6 environments in PPAR γ and similar environments detected in enzyme structures. The PPAR γ environment defined around fragment 5 (Figure 7a) was described by 11 surface feature points, 7 of which could be matched to a phenyl environment found in the active site of phospholipase A2 (EC 3.1.1.4). From a total of 71,299 enzyme environments screened, this solution was found in rank 2 among the 183 unique chemical fragments mapped to this environment. An amide group occupied rank 1, so rank 2 was in fact the first solution with a ring-containing fragment. Despite interacting with a phenyl group, the PPAR γ environment surrounding fragment 7 is essentially different from that of fragment 5 (Figure 7b). In this case, a solution in rank 10 identified a phenyl environment present in the binding cavity of 5-methyltetrahydrofolate corrinoid/iron sulfur protein methyltransferase (EC 2.1.1.13). Among the top nine positions, there were five copies of a propyl chain, two copies of an ethyl chain, one copy of a dimethylsulfide chain, and one copy of a cyclohexyl ring in rank 8. The phenyl solution in rank 10 is thus in fact the fifth solution among the 30 unique chemical fragments mapped to this environment, the second one containing a ring. The PPAR γ environment surrounding fragment 8 was defined by nine surface feature points (Figure 7c).

Table 1. List of Top-Scoring Ring-Containing Fragments with at Least One Heteroatom and at Most One Chiral Center from Enzyme Protein Environments Mapped onto the Environments of a PPAR γ Structure (pdb entry 1fm9)^a

Fragment	S	N	N _{PDBbind}	%Public	N _{Actives}	Representative PPAR γ active
1 	0.75	1	14	0.20	12	
2 	0.74	1	5	0.06	5	
3 	0.71	2	83	3.18	21	
4 	0.71	4	16	0.48	12	
5 	0.70	1	43	0.20	0	-
6 	0.70	7	18	10.27	659	

^aAlso added is the similarity score (S), number of copies mapped (N), number of instances in PDBbind (N_{PDBbind}), percentage of bioactive compounds from public sources^{37–39} containing the fragment (%Public) and number (N_{Actives}), and a representative structure of bioactive PPAR γ compounds found in those public sources.^{37–39}

A total of 52 unique chemical fragments were mapped to this PPAR γ environment directly from unrelated enzyme environments. The top solution contained a carboxylate group originally found in [ribulose-bisphosphate carboxylase]-lysine N-methyl-transferase (EC 2.1.1.127), very much in agreement with the functional group found in the native PPAR γ ligand. A total of nine surface feature points described also the environment in PPAR γ linked to the phenyl ring in fragment 9 (Figure 7d). Of those, six were found to match closely to a phenyl protein environment in Rous sarcoma virus retropepsin (EC 3.4.23.B10), the best solution found among the list of 110 chemical fragments mapped to this environment. Fragment 10 is a small remaining portion of an aliphatic linker chain of the ligand that interacts deep enough in the binding cavity of PPAR γ to have a surrounding environment described by seven surface feature points (Figure 7e), five of which superimpose optimally with an ethylene environment from a HIV-1 retropepsin structure (EC 3.4.23.16) occupying rank 4 from a total of 193 fragments mapped. Finally, 11 surface feature points define the PPAR γ environment that interacts with the phenyl ring of fragment 11 (Figure 7f), of which six can be optimally matched to a phenyl environment located in the vascular endothelial growth factor

receptor 2 kinase (EC 2.7.1.112). Among the 132 chemical fragments mapped directly into this PPAR γ environment, a phenyl ring emerges as the second best solution.

It is important to stress that among all chemoisosteric relationships identified between pairs of environments in PPAR γ and unrelated enzymes (including three transferases and three hydrolases), six out of the nine chemical fragments extracted from the native ligand (Figure 6b) could be recovered in the right location and orientation among the top solutions without any further computational refinement. Depending on the size and complexity of the complete binding cavities, some of these chemoisosteric relationships may actually translate into poly-target pharmacology for some small molecules. As an example, indomethacin is a nonsteroidal noninflammatory agent that has been reported to show affinity for both PPAR γ and PLA2,³⁷ which is precisely one of the local chemoisosteric relationships identified (Figure 7a). This illustrates nicely the potential use of chemoisosterism to guide medicinal chemistry efforts in fragment-based lead generation, both with respect to suggesting the right vectors and fragments for optimal structure progression and to alerting of likely undesired off-target affinities.

Are Other Fragments Present in Bioactive Small Molecules?

Another aspect of fragment-based drug discovery in which the concept of chemoisosterism can be widely applied is in the design of fragment libraries for focused or targeted screening.⁴⁶ Having shown in the previous section that the approach is able to reconstruct the main interacting chemical fragments of the native ligand in a PPAR γ structure, one could extrapolate these findings to all fragments projected inside the PPAR γ binding cavity (Figure 6d). This is equivalent to assume that if those chemoisosteric relationships could be confirmed, chances are that the same may hold true as well for all other enzyme environments found similar to PPAR γ environments. In total, among the 8893 unique chemical fragments found in enzyme environments that have at least one ring, at least one heteroatom, and at most one chiral center, a selected set of 134 fragments from 302 enzyme environments matching a PPAR γ environment (similarity score ≥ 0.60) were placed in the binding cavity of that PPAR γ structure (PDB code 1fm9). This list of fragments could constitute the seed for a focused fragment library designed to probe the PPAR γ binding cavity. Indeed, 65 of them were ultimately found in bioactive molecules to PPAR γ . The top-scoring fragments (similarity score ≥ 0.70) are collected in Table 1.

As a means to assess whether each of the chemical fragments proposed could be amenable to interact favorably with PPAR γ , substructure searches were performed on public chemistry sources of pharmacological data^{37–39} to confirm the presence of some of those fragments in small molecules with affinity (pK_i , pIC_{50} or $pEC_{50} > 6$) to PPAR γ . As observed, known bioactive molecules to PPAR γ were identified for the majority of the chemical fragments listed in Table 1. For example, a methoxycyclopentyl fragment (**1**) was identified in the structures of 12 bioactive molecules to PPAR γ , a representative of which ($pIC_{50} = 7.2$) is included also in Table 1. Among all fragments in Table 1, the most represented in the structures of PPAR γ bioactive molecules is, with 659 examples, a methoxyphenyl fragment (**6**), whereas the least represented is a 1,2,3-triazoline fragment (**2**), with only 5 PPAR γ bioactive molecules found. Interestingly, the relative presence of these fragments in PPAR γ bioactive compounds agrees qualitatively with the rareness of the same fragments in general bioactive molecules present in public sources.^{37–39} This highlights the fact that the present approach is able to retrieve chemical fragments that are compatible with the interaction to a particular protein environment, not necessarily implying they should be considered privileged fragments. Special mention should be made of the presence among this top-scoring chemical fragments of a pyrimidine-2,4(1H,3H)-dione fragment (**5**). This fragment was not found in the structure of any of the bioactive molecules to PPAR γ currently present in public chemistry sources with pharmacological data. However, it is closely related to the thiazolidinedione group present in many PPAR γ bioactive compounds (see the representative PPAR γ active for fragment **1**), including some drugs such as troglitazone and rosiglitazone.

Further to validating the mere presence of the fragments listed in Table 1 in the structures of PPAR γ bioactive compounds, one should check whether the positions and orientations of those fragments in the binding site of PPAR γ match with the geometries observed in the corresponding fragment of co-crystallized PPAR γ ligands. To address this point, all PPAR γ X-ray structures available in the PDB were retrieved (122 structures in total). Of those, 75 structures contained a ligand occupying the region of the binding site defined by our reference PPAR γ (PDB code 1fm9). The geometries of the chemical fragments extracted from

those 75 ligands in the binding site of PPAR γ constitute a good template for validation. Among the six fragments listed in Table 1, the methoxyphenyl fragment was found present in four of the co-crystallized ligands. Therefore, despite the limited amount of data, a comparison can be performed between the geometries of the four methoxyphenyl fragments as observed in the respective crystal structures and the positions and orientations obtained for the seven methoxyphenyl fragments mapped directly to the PPAR γ binding site from chemoisosteric enzyme environments. The results are shown in Figure 8. As observed, even though

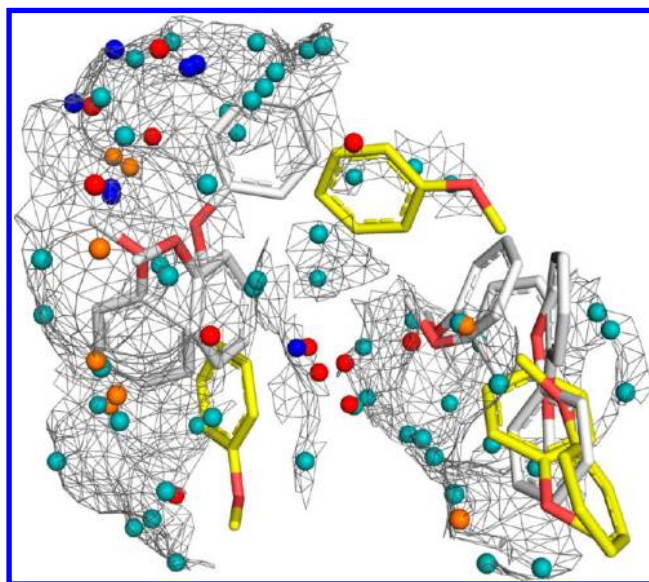


Figure 8. Relative positions and orientations of the seven methoxyphenyl fragments found in chemoisosteric enzyme environments (white carbon atoms) directly mapped in the binding site of PPAR γ compared to the geometries of the four methoxyphenyl fragments present in ligands co-crystallized into the PPAR γ binding site (yellow carbon atoms).

matches are not perfect, fragments tend to cluster in similar regions of the binding site.

One could find at least three reasons that may contribute to this dispersion. First, there is a huge lack of completeness in protein structure data, and the fact that the interaction in that geometry is not observed does not imply it may not be compatible. Second, the fragments extracted from co-crystallized ligands (yellow carbon atoms) are all accommodated in slightly different PPAR γ binding sites, which differ also from the one used to map the chemical fragments from chemoisosteric relationships (white carbon atoms). Third, all seven chemical fragment geometries were mapped directly from the best clique solution found with the chemoisosteric enzyme environments, and no interaction energy optimization was applied to allow their positions and orientations to relax in the binding site of PPAR γ . Taking all this into consideration, these results confirm that chemoisosterism can be used to design fragment libraries directed to particular protein targets or target families.

Inferring Chemoisosterism without Similarity Constraints. Even though looking for protein environments that expose similar pharmacophoric features may seem as the most natural way of inferring chemoisosteric relationships to generate a priority list of fragments likely to bind to a given protein cavity (as shown in Figure 6), the whole process does not need to be limited to the identification of cavities bound to a certain degree

of similarity and constrained by the ability of any approach to detect similar interacting patterns in those cavities. In this respect, the construction of fragment-environment interaction networks derived from structure-based information on protein–ligand complexes, such as those shown in Figures 3 and 5a, offers a means to infer additional relationships directly based on existing connections and that go beyond any similarity constraints both between protein environments and chemical fragments.

Network guilt-by-association is a well-established approach for inferring missing links between concepts in bipartite networks, as recently applied to predict gene-phenotype and drug-side effect associations, among others.^{47,48} The basic assumption is that protein environments linked to a common chemical fragment will likely share other chemical fragments as well. For the present case, Figure 9 illustrates the process applied to the adenyl-

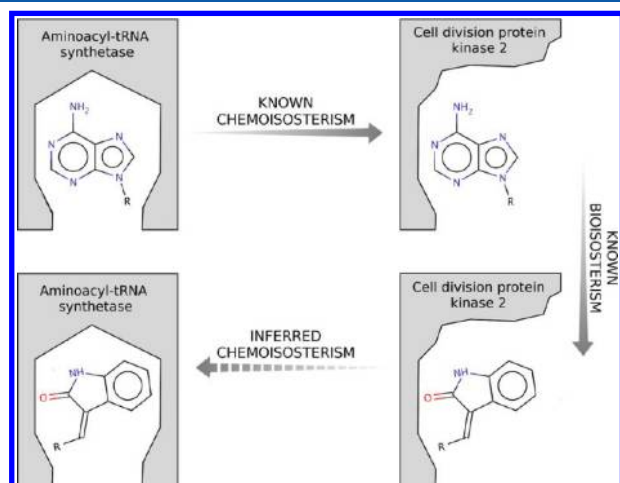


Figure 9. Process of inferring bioisosterism from chemoisosterism: (a) starting from a structure of a protein–ligand complex, (b) chemoisosteric environments are obtained, and (c) bioisosteric fragments to all those environments are evaluated as possible bioisosteres to the starting fragment. (d) A surface overlapping score is used to rank the fragments. The scheme continues with the same examples from Figure 1.

centered network included in Figure 5a. As a reminder, in the network of Figure 5a, the first layer of colored dots represents all protein environments, similar or not, known to interact with the adenyl fragment and thus related by chemoisosterism. Therefore, as schematically represented in Figure 9, structural evidence

exists that environments in the binding cavity of aminoacyl-tRNA synthetase (aa-tRNA; PDB code 2v0c, EC 6.1.1.4, depicted as a large green dot in the first layer of Figure 5a) and cell division protein kinase 2 (CDK2; PDB code 1b38, EC 2.7.11.22, depicted as large yellow dot in the second layer of Figure 5a) are related by chemoisosterism to the adenyl fragment. But in addition, structural evidence exists also that the same CDK2 environment binds to a wide diversity of fragments (depicted as black dots in a third layer connected to the large yellow dot in the second layer of Figure 5a). One of those is the 3-methylideneoxindole fragment (PDB code 1ke9), and it is thus bioisosteric to the adenyl fragment in that environment of the CDK2 binding cavity. Therefore, if the aa-tRNA and CDK2 environments are chemoisosteric relative to the adenyl fragment and the adenyl and 3-methylideneoxindole fragments are bioisosteric relative to the CDK2 environment, one may infer that the aa-tRNA and CDK2 environments are also chemoisosteric relative to the 3-methylideneoxindole fragment and thus that the adenyl and 3-methylideneoxindole fragments are also bioisosteric relative to the aa-tRNA environment.

Using the fragment-environment interaction network displayed in Figure 3, the process outlined in Figure 8 was applied to the same complex of GI262570 and PPAR γ used previously. Because the network in Figure 3 is built from fragments composed of a single ring system, only the phenyl and methyl oxazole fragments in GI262570 could be queried against the network (Figure 6b). In total, 167 and 7 chemoisosteric relationships were inferred using the phenyl and methyl oxazole fragments, respectively, resulting in a final list of 192 unique fragments composed of a single ring system and containing at least one heteroatom and at most one chiral center. Among those, 67 fragments were found to be present in bioactive small molecules for PPAR γ .

Comparison with Other Strategies. The term chemoisosterism among protein environments has been introduced as a natural complementary concept to the widely established notion of bioisosterism among chemical fragments, and two implementations of the concept have been applied to identify a list of fragments composed of a single ring system to probe the binding cavity of PPAR γ using structural information from distantly related proteins only. At this stage, one may argue whether those directed fragment selections are any different from what one would get using more established approaches for this task, such as ligand similarity and protein–ligand docking. Accordingly, we used pharmacophoric fragment (PHRAG) descriptors⁴⁹ to rank order chemical fragments based on their similarity against all

Table 2. Number of Unique Chemical Fragments Found in Bioactive Molecules to PPAR γ (positives) in Both Set1 and Set2 (see text for definitions), Total Number of Fragments Selected from Chemoisosterism and Retrieved as Top Scoring from Similarity and Docking Predicted To Bind to PPAR γ (predictive positives), and Number of Those among Positives (true positives).^a

method	fragment set	positives	predicted positives	true positives	precision	recall	F-measure
chemoisosterism clique	Set1	700	134	65	0.49	0.09	0.16
ligand similarity PHRAG (top 134)	Set1	700	134	59	0.44	0.08	0.14
ligand–protein docking AutoDock (top 134)	Set1	700	134	39	0.29	0.06	0.09
union unique	Set1	700	355	134	0.38	0.19	0.25
chemoisosterism network	Set2	109	192	67	0.35	0.61	0.45
ligand similarity PHRAG (top 192)	Set2	109	192	59	0.31	0.54	0.39
ligand–protein docking AutoDock (top 192)	Set2	109	192	64	0.33	0.59	0.42
union unique	Set2	109	389	98	0.25	0.90	0.39

^a Also added are the levels of precision, recall, and corresponding F-measure obtained from the different approaches, including their union, obtained when considering the fragments retrieved by all approaches.

fragments extracted from GI262570 and AutoDock⁴² to prioritize chemical fragments based on their goodness of fit (predicted binding energy or ligand efficiency) inside the binding cavity of PPAR γ (PDB code 1fm9). PHRAGs are defined by all possible fixed-length segments of five atom-features that can be extracted from the topology of a molecule.⁴⁹ These approaches were applied to two fragment sets: the set of 8893 unique chemical fragments (Set1) extracted from ligands present in enzyme entries in PDBbind containing at least one ring, at least one heteroatom, and at most one chiral center, used to infer chemoisosterism under similarity constraints (clique chemoisosterism), and a derived subset of 731 chemical fragments (Set2) composed of a single ring system used to infer chemoisosterism without similarity constraints (network chemoisosterism). Set1 and Set2 were found to contain 700 and 109 fragments present in the structures of known bioactive compounds to PPAR γ , respectively.

Table 2 summarizes the results obtained. For Set1, clique chemoisosterism returned 134 fragments, from which 65 were found within the list of 700 fragments in Set1 present in known PPAR γ bioactive ligands (vide supra). This gives a level of recall (true positives/positives; also referred to as sensitivity or false positive rate) of 9.3% and a level of precision (true positives/predicted positives; also referred to as positive predictive value) of 48.5%. Comparatively, 59 and 39 true positives were identified within the top-scoring 134 fragments using PHRAG similarity and AutoDock ligand efficiency, respectively. For Set2, network chemoisosterism returned 192 fragments, from which 67 were found within the list of 109 fragments in Set2 presently in known PPAR γ bioactive ligands (vide supra). This corresponds to levels of recall and precision of 61.5% and 34.9%, respectively. Comparatively, 59 and 64 true positives were identified within the top-scoring 192 fragments using PHRAG similarity and AutoDock ligand efficiency, respectively.

In spite of the relatively similar levels of precision and recall found among the different approaches, one may argue though that the diversity of the chemical fragments selected from the docking approach may be significantly higher than the diversity of fragments selected from chemoisosterism and specially the similarity approach but also that the fragment selection obtained from chemoisosterism may not add any value to selections derived from similarity and docking and only be a subset of them. To address these two points, diversity and complementarity, an all-against-all similarity analysis of the chemical fragments selected from each approach was performed using PHRAG descriptors. Figure 10 displays the kernel densities of similarity values obtained when comparing the true positive fragments obtained from chemoisosterism, similarity, and docking on Set1 (Table 2). It is not surprising that fragments retrieved by similarity are intrinsically more similar than those selected by the other two approaches, but it is most interesting to realize that fragments identified by chemoisosterism and docking have a similar diversity profile. In addition, a Venn diagram is also included in Figure 10 to illustrate the relative overlap among the fragment sets selected by the different methods. Remarkably, there is a striking level of complementarity, with only three fragments being shared between the three approaches. Therefore, fragments retrieved by chemoisosterism are not only diverse among them but they are also complementary to those found by similarity and docking approaches. This complementarity is reflected in Table 2. Taking all unique fragments resulting from the union of selections obtained by the different approaches leads to a significant gain in recall compared to any individual

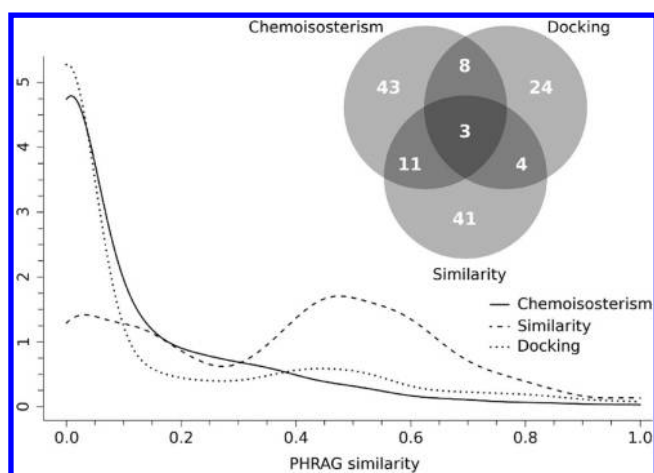


Figure 10. Kernel densities for intraset similarities for the fragments retrieved from Set1 using chemoisosterism with clique detection, similarity with PHRAG descriptors, and docking with AutoDock and ligand efficiency scoring. A Venn diagram illustrating the overlap between the true positive fragments identified by each approach is also included.

approach, without penalizing precision strongly, in both Set1 and Set2. Obviously, strictly different results may be obtained by using different descriptors for assessing ligand similarity and different sampling and scoring implementations for fragment docking,^{50,51} but the results presented in Table 2 and the analyses shown in Figure 10 are certainly indicative of the direct applicability of chemoisosterism for fragment library design compared to other well-established approaches.

CONCLUSIONS

The current perception of chemoisosterism is bound to some inherent limitations related to the actual coverage and bias of both chemical fragments and protein environments in the structural data available on protein–ligand complexes. In terms of coverage, approximately 34% of the ring-containing chemical fragments that can be extracted from drugs³⁷ are represented in the structures of ligands contained in PDBbind, and almost 80% of all entries in PDBbind correspond to enzyme structures. Therefore, increasing the coverage of biologically relevant chemical fragments present in ligands co-crystallized with proteins, but also the number of representative structures from members of therapeutically relevant protein families other than enzymes, is an absolute requirement to widen the scope of applicability of chemoisosteric searches. In terms of bias, basically about 20% of chemical fragments extracted from all ligand structures contained in PDBbind are linked to 80% of protein environments. Nevertheless, the proportion ranking of each individual chemical fragment in PDBbind was found to be relatively comparable to that derived from all drugs in DrugBank (FDA-approved small molecule drugs plus experimental drugs)³⁷ and all bioactive molecules in other public sources (vide supra for some of the chemoisosteric hotspots identified).^{37–39} In this respect, the fragment bias observed in PDBbind is not a particular feature of this source but indicates some general trends in public sources of bioactive molecules.^{37–39} It has been long a matter of debate to what extent this may reflect some degree of biological bias in medicinal chemistry or is simply due to a pure chemistry bias based on synthetic accessibility.

In spite of these limitations, an analysis of the interactions between chemical fragments and protein environments in currently available structures of protein–ligand complexes has revealed the existence of local chemoisosteric relationships among the binding cavities of completely unrelated proteins. In this respect, it is particularly revealing that one can obtain a reasonably correct fragment mapping of the binding cavity of a nuclear receptor structure by direct superposition of protein environments from enzyme structures only and without any further computational energy optimization to refine the placement and orientation of the associated chemical fragments. In addition, it has been also shown that one can exploit fragment–environment interaction networks for designing fragment libraries to probe a certain protein cavity at virtually no computational cost and comparable performance to other well-established methods.

Beyond this particular application, the extent by which chemoisosterism is linked to all constitutional fragments of a chemical structure has multiple implications for medicinal chemistry efforts in fragment-based drug discovery. Provided that the structure of the target protein is available, the identification of chemoisosteric relationships among protein environments emerges as a natural concept, fully complementary with the widely accepted notion of bioisosterism, that can be directly applied from designing targeted libraries for fragment screening, through guiding the structural evolution of an original fragment hit, up to scaffold hopping and de novo design of novel chemical series.

■ ASSOCIATED CONTENT

■ Supporting Information

Bioisosteric hotspots defined by a set of 41 thrombin environments (Figure S1) and a set of 117 protein environments from 7 different serine proteases (Figure S1). This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jmestres@imim.es. Tel: +34 93 3160540. Fax: +34 93 3160550.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by the Spanish Ministerio de Ciencia y Educación (BIO2011-26669 and PTA2009-1865-P) and the Instituto de Salud Carlos III.

■ REFERENCES

- (1) Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: Redesign web site and web services. *Nucleic Acids Res.* **2011**, *39*, D392–401.
- (2) Mestres, J. Representativity of target families in the Protein Data Bank: Impact for family-directed structure-based drug discovery. *Drug Discovery Today* **2005**, *10*, 1629–37.
- (3) Weigelt, J. Structural genomics-impact on biomedicine and drug discovery. *Exp. Cell Res.* **2010**, *316*, 1332–1338.
- (4) Hubbard, R. E. Structure-based drug discovery and protein targets in the CNS. *Neuropharmacology* **2011**, *60*, 7–23.
- (5) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
- (6) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: A library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.
- (7) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttmann, C. Identification of target-specific bioisosteric fragments from ligand–protein crystallographic data. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 385–394.
- (8) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inf.* **2010**, *29*, 366–385.
- (9) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
- (10) Southall, N. T. Ajay kinase patent space visualization using chemical replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.
- (11) Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **2011**, *54*, 2529–2591.
- (12) 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.
- (13) Briansó, F.; Carrascosa, M. C.; Oprea, T. I.; Mestres, J. Cross-pharmacology analysis of G protein-coupled receptors. *Curr. Top. Med. Chem.* **2011**, *11*, 1956–1963.
- (14) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.
- (15) Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.
- (16) Garcia-Serna, R.; Mestres, J. Anticipating drug side effects by comparative pharmacology. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 1253–1263.
- (17) Wang, X.; Greene, N. Comparing measures of promiscuity and exploring their relationship to toxicity. *Mol. Inf.* **2012**, *31*, 145–159.
- (18) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
- (19) Wagener, M.; Lommerse, J. P. M. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.
- (20) Zauhar, R. J. SMART: A solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 149–159.
- (21) 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.
- (22) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607–633.
- (23) Gold, N. D.; Jackson, R. M. SitesBase: A database for structure-based protein–ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34*, D231–4.
- (24) Kufareva, I.; Ilatovskiy, A. V.; Abagyan, R. Pocketome: An encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res.* **2012**, *40*, D535–540.
- (25) Ito, J.-I.; Tabei, Y.; Shimizu, K.; Tsuda, K.; Tomii, K. PoSSuM: A database of similar protein–ligand binding and putative pockets. *Nucleic Acids Res.* **2012**, *40*, D541–548.
- (26) Niskanen, S.; Östergård, P. R. J. *Cliques User's Guide*, Version 1.0; Communications Laboratory, Helsinki University of Technology: Espoo, Finland, 2003.
- (27) Mestres, J.; Maggiora, G. Putting molecular similarity into context: Asymmetric indices for field-based similarity measures. *J. Math. Chem.* **2006**, *39*, 107–118.
- (28) Gregori-Puigjané, E.; Garriga-Sust, R.; Mestres, J. Indexing molecules with chemical graph identifiers. *J. Comput. Chem.* **2011**, *32*, 2638–2646.
- (29) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.
- (30) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, 2009; pp 485–586.

- (31) Vogt, I.; Mestres, J. Drug-Target Networks. *Mol. Inf.* **2010**, *29*, 10–14.
- (32) 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.
- (33) Edfeldt, F. N. B.; Folmer, R. H. A.; Breeze, A. L. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today* **2011**, *16*, 284–287.
- (34) Surade, S.; Blundell, T. L. Structural biology and drug discovery of difficult targets: The limits of ligandability. *Chem. Biol.* **2012**, *19*, 42–50.
- (35) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **2009**, *27*, 157–167.
- (36) Liao, J. J.-L. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J. Med. Chem.* **2007**, *50*, 409–424.
- (37) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for “omics” research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–1041.
- (38) Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discovery* **2011**, *10*, 197–208.
- (39) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15*, 1052–1057.
- (40) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: New pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550–557.
- (41) Kalinina, O. V.; Wichmann, O.; Apic, G.; Russell, R. B. Combinations of protein–chemical complex structures reveal new targets for established drugs. *PLoS Comput. Biol.* **2011**, *7*, e1002043.
- (42) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- (43) Schubert, C. R.; Stultz, C. M. The multi-copy simultaneous search methodology: A fundamental tool for structure-based drug design. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 475–489.
- (44) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187–192.
- (45) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.
- (46) Chessari, G.; Woodhead, A. J. From fragment to clinical candidate—A historical perspective. *Drug Discovery Today* **2009**, *14*, 668–675.
- (47) Yang, P.; Li, X.; Wu, M.; Kwok, C.-K.; Ng, S.-K. Inferring gene-phenotype associations via global protein complex network propagation. *PLoS ONE* **2011**, *6*, e21502.
- (48) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.
- (49) Vidal, D.; Garcia-Serna, R.; Mestres, J. Ligand-based approaches to in silico pharmacology. *Methods Mol. Biol.* **2011**, *672*, 489–502.
- (50) Kawatkar, S.; Wang, H.; Czerminski, R.; Joseph-McCarthy, D. Virtual fragment screening: An exploration of various docking and scoring protocols for fragments using Glide. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 527–539.
- (51) Englert, L.; Silber, K.; Steuber, H.; Brass, S.; Over, B.; Gerber, H.-D.; Heine, A.; Diederich, W. E.; Klebe, G. Fragment-based lead discovery: Screening and optimizing fragments for thermolysin inhibition. *ChemMedChem* **2010**, *5*, 930–940.