JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

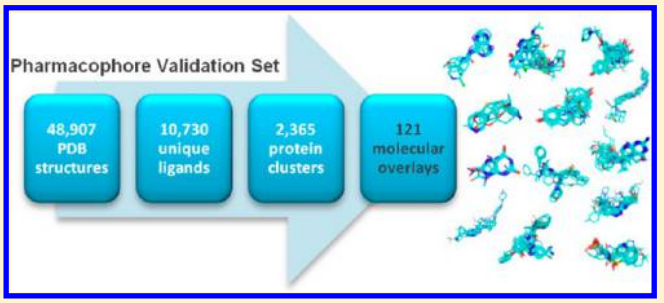
An Extensive and Diverse Set of Molecular Overlays for the Validation of Pharmacophore Programs

Ilenia Giangreco,* David A. Cosgrove, and Martin J. Packer

AstraZeneca, Mereside, Alderley Park, Macclesfield SK10 4TG, United Kingdom

Supporting Information

ABSTRACT: The pharmacophore hypothesis plays a central role in both the design and optimization of drug-like ligands. Pharmacophore patterns are invoked to explain the binding affinity of ligands and to enable the design of chemically distinct scaffolds that show affinity for a protein target of interest. The importance of pharmacophores in rationalizing ligand affinity has led to numerous algorithms that seek to overlay ligands based on their pharmacophoric features. All such algorithms must be validated with respect to known ligand overlays, usually by extracting ligand overlay sets from the Protein Data Bank (PDB). This validation step creates the problem of which of the known overlays to select and from which proteins. The large number of structures and protein families in the PDB makes it difficult to establish a definitive overlay set; as a result, validation studies have rarely employed the same data sets. We have therefore undertaken an exhaustive analysis of the RCSB PDB to identify 121 distinct ligand overlay sets. We have defined a robust protein overlay protocol, which is free from subjective interpretation over which residues to include, and we have analyzed each overlay set on the basis of whether they provide evidence for the pharmacophore hypothesis. Our final data set spans a broad range of structural types and degrees of difficulty and includes overlays that any algorithm should be able to reproduce, as well as some for which there is very weak evidence for a conserved pharmacophore at all. We provide this set in the hope that it will prove definitive, at least until the PDB is greatly enriched with further structures or with radically different protein folds and families. Upon publication, the data set will be available for free download from the Web site of the Cambridge Crystallographic Data Centre.



INTRODUCTION

Three-dimensional computer-aided molecular design is frequently divided into two strategies, structure-based design (SBD) and ligand-based design (LBD). The former requires the availability of a 3D structure of the macromolecule (enzyme or receptor) for which one is attempting to find a tightly binding ligand; the latter is most often used in those cases where such a structure does not exist. Examples of techniques of use in LBD are pharmacophore analysis[1] and 3D quantitative structure activity relationships (3D QSAR),[2] both of which require as input a set of molecules with known affinity to the macromolecule of interest.

The first, and frequently most challenging, step in both pharmacophore analysis and 3D QSAR is a superimposition of one low energy conformer of each molecule in a manner that it is hoped mimics the relative configurations of the ligands when bound in the active site of the target.[3] Normally, this is achieved by generating an ensemble of low energy conformations of each ligand (either as a preliminary step or as part of the overlay process), selecting one conformation of each, and overlaying the selected conformations such that one or more scoring functions are optimized. Over the decades that computer-aided molecular design has existed as a subject, there have been numerous different methods developed both for performing and scoring the overlays. Leach et al. have recently reviewed this in some detail.[4] Nevertheless, these methods are still being developed with enthusiasm, and this is a clear indication that the overlay problem is not yet adequately solved.[5,6]

One common feature of all publications of new overlay procedures is the need to demonstrate that they work in at least some situations. This is almost always done by reference to structures drawn from the Protein Data Bank (PDB). The procedure generally follows the sequence:

1. Select an enzyme for which multiple structures are available with different ligands in the active site.
2. Overlay the proteins into a common reference frame based on some set of features common to all of them (backbone atoms or alpha carbons, for example, of either all residues in the protein, or a subset selected by some means).
3. Extract the ligands from the protein structure in the common reference frame and denote this the experimental overlay.
4. Use the overlay program of interest to superimpose the ligands.

Received: January 11, 2013
Published: April 8, 2013

   dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

5. Compare the results of the overlay with the experimental overlay, either by visual inspection or by calculating some parameter such as the root-mean-squared distance (RMSD) between corresponding atoms in the theoretical and experimental overlays.

Generally, the number of enzyme/ligand structure sets used in a particular paper is small. In addition, it is unusual for the authors of two different overlay programs to report tests on the same enzymes, and even when they do, it is not uncommon for them to select different subsets of ligands binding to the same enzyme and/or overlay the enzymes by different protocols. For example, in 2002, Patel et al.[7] compared three commercially available programs (Catalyst/HipHop,[8] GASP,[9] and DISCO[10]) by generating known pharmacophores from five different protein families: thrombin, cyclin dependent kinase 2 (CDK2), dihydrofolate reductase (DHFR), HIV reverse transcriptase, and thermolysin. For each of the five proteins, a number of PDB complexes were overlaid and visually inspected to identify a target pharmacophore that was used as a reference to estimate the accuracy of the three programs. In 2006, the same data set was used to validate PHASE.[11,12] In the same year, Richmond at al. validated GALAHAD,[13] a pharmacophore program released with SYBYL 7.2, and added to the five targets cited above a set of CDK-2 inhibitors compiled by Thierry Langer for the Fifth European Workshop on Drug Design. The version of FLAME[14] able to generate multiple flexible alignments was tested on five structurally diverse D3 receptor ligands downloaded from the PDB. In 2010, Jones validated his algorithm GAPE[15] with 13 sets of protein−ligand complexes extracted from the PDB. Ten of these sets from 7 different targets (CDK2, elastase, ESR1, HIV-1 protease, p38, rhinovirus, and trypsin) were chosen from a previous work,[16] which examined multiple factors influencing the quality of overlays obtained through ROCS and FlexS. In addition, he used one set of DHFR complexes and two sets of Factor Xa (FXa) complexes, all retrieved from the PDB. The DHFR set is a superset of the corresponding set used by Patel et al.[7] A subset of six of these data sets has been used by Klabunde et al.[5] for the validation of MARS (multiple alignments by ROCS-based similarity). In 2012, Taylor et al.[17] developed a program for overlaying multiple flexible molecules and tested it on 10 sets of protein−ligand complexes from the PDB, members of the Astex Non-Native Set.[18] Recently, Cross et al. validated FLAPpharm,[6] a GRID-based pharmacophore elucidation approach, with 81 sets of overlaid ligand structures, including the five targets from the Patel data set.[19] This large data set, called PharmBench, was produced to address a gap in the area of pharmacophore validation where a benchmarking set of molecular overlays was not available, and therefore, it was difficult to compare objectively the performance of one program versus another.

In this paper, we report the development of a protocol for selecting sets of protein−ligand complexes suitable for use in the validation of overlay procedures, placing the proteins in each set into a common reference frame and extracting the ligands to produce the final validation sets. We have attempted to make this protocol as automated as possible to reduce subjectivity in the selection of protein−ligand complexes, the selection of residues used to align the protein structures, and finally, in those enzyme classes where there is a large set of structures, the selection of an appropriately sized smaller subset. This work started before the publication of PharmBench with the aim of addressing the same gap in the literature.

Similarly to the well-curated set of protein−ligand structures used to validate docking programs (the so-called Astex Diverse Set),[20] the Cambridge Crystallographic Data Centre (CCDC) has also generously agreed to host this overlaid set of structures. In doing so, our hope, and the CCDC's, is to free developers of novel overlay programs from the additional burden of having to generate their own validation sets, not least because this burden is sometimes used as a reason to publish relatively small validation studies, and there is occasionally a suspicion that the small set selected has been chosen to show the overlay program in the best possible light. In the future, they will have easy access to a large and diverse array of sets of overlaid ligands with which to demonstrate the value of their contribution to the field.

We believe that this data set is complementary to PharmBench in many respects. In the latter, targets were selected as being those for which an approved drug is a ligand, for which structures are available, and electron densities deposited, although the authors do not state whether the approved drugs are always in the final set. We have used different criteria for target selection and, thus, obtained a more comprehensive set both in number of targets and number of ligands, with a more precise method of overlaying the protein structures from each target class. More specifically, of the 121 sets of overlaid molecules, from 119 targets, for which we provide an overlay, only 33 appear in the paper of Cross et al.

This paper describes the protocols used to select the protein−ligand complexes, the method of selecting the residues used to overlay each different set of complexes, and the rules for selecting a subset of ligands where appropriate. A full list of the structures used is given in the Supporting Information.

MATERIALS AND METHODS

Selection of Complexes. We have derived a pharmacophore validation set using structures available in the public domain, specifically modern and high-quality protein−ligand complexes from the RCSB PDB. To this aim, we limited the selection to X-ray structures with a resolution of 2.5 Å or better, determined between January 2000 and May 2012, when we started this analysis. The limit on the date aims to exclude structures determined and refined without modern crystallographic procedures.

While we acknowledge the importance of inspecting electron densities when analyzing bound ligands,[20] we purposely did not restrain our search on the basis of structure factors. The overlay process that we outline below creates ligand sets that should have some pharmacophore relationship. If these sets contain artifacts or ligands that are poorly fit to their density, then the overlays will rank low based on our analysis of coincident features, and we therefore flag them as difficult to reproduce.

On the basis of these preliminary selection criteria, we found 48,907 protein−ligand complexes. Using the PDB, the user is allowed to save the search results in a customized table collecting a wide range of information. In this work, we selected a few fields of interest, mostly relating to biological details of the protein (e.g., EC number, source of expression) and ligand details (e.g., formula, SMILES string, identifier). The table contained 217,389 rows, one for each non-protein entity (e.g., ligand, ion, cofactor) found in each PDB entry.

However, the table contained much redundant information that needed removing to give a clean data set with which to

work. We followed a hierarchical filtering process described in the Ligand Analysis section (Table 1).

Table 1. Rules To Filter the Initial Set of Data

filters	no.[a]
exclude compounds with ligand ID null	189,982
exclude solvents and small ions[b]	91,999
exclude cofactors[b]	83,394
exclude compounds with HAC ≤ 10	53,027
exclude compounds with nonorganic elements	51,861
exclude simple saccharides	45,159
exclude compounds with $(CH_2)_4$, or $CH_2O(CH_2CH_2O)_2$, linkers	38,870
exclude compounds that fail the rule of five	22,299
exclude compounds that occur more than 20 times	10,730

[a]Numbers of ligands surviving each filter. [b]List taken from *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.

Ligand Analysis. The three letter code that uniquely identifies a ligand structure in the PDB was a key detail in the filtering process. Therefore, we first excluded those rows in the table where there was no ligand. Then, we removed solvents, small ions, cofactors, and simple saccharides. In addition, using the SMILES string of each ligand supplied by the PDB, we calculated a number of physical properties that are necessary to pick up only drug-like molecules. These physical properties included heavy atom count, molecular weight, ClogP,[21] number of hydrogen bond donors, and number of hydrogen bond acceptors. This excluded ligands with less than 10 heavy atoms and ligands that fail Lipinski's rule of five.[22] Finally, we defined SMARTS strings to select only ligands with standard organic atoms (i.e., C, N, O, P, S, Br, Cl, and F) and exclude polyethers or highly flexible molecules with more than four consecutive methylene linkers. As a final filter, we excluded ligands that occurred more than 20 times because we assumed they are not interesting molecules for pharmacophore purposes. Examples include buffering agents (e.g., MES), capping groups for amino acids (e.g., ACE), and post-translationally modified amino acids (e.g., SEP, TYS).

We deliberately decided not to use experimental affinity data, even where it may be available. It is not rational to suppose that pharmacophores can explain affinity differences with any accuracy, given the range of factors that dictate an observed overlay, beyond that of the pharmacophore model. It would not be difficult to obtain this experimental data from public databases, in those few cases where it is extant.

Protein Grouping. Once we obtained a "clean" set of ligands, we moved on to classify the proteins. First, we used the EC number for a top level distinction of the enzymes (i.e., oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases) from other types of structures such as receptors. Then, we used the UniProt ID[23] for a more focused grouping of proteins; this allowed the distinction of proteins expressed from the same gene but from different species.

It is worth stressing that we aimed at deriving a high quality diverse set of ligands useful for pharmacophore elucidations. With this in mind, we selected only unique ligands within each group of structures, and where multiple structures of the same protein−ligand complex were available, we chose the structure with the best resolution. The classification herein described yielded 2365 clusters, but we followed up only 183 clusters that contained at least 5 ligands. These clusters were distributed as follows: 33 oxidoreductases (EC 1), 55 transferases (EC 2), 58 hydrolases (EC 3), 4 lyases (EC 4), 4 isomerases (EC 5), 3 ligases (EC 6), and 26 non-enzyme proteins. A detailed list of all targets is provided in Table S1 of the Supporting Information.

The protocol described thus far was fully automated; the remainder of the process requires some manual intervention.

Overlaying Complexes. Central to the pharmacophore concept is the overlay of ligands, which is used as a starting point to identify interactions with the protein that they have in common. In this work, we attempted a supervised analysis by exploiting the 3D structural information of a high number of protein targets.

The basic assumption is that given a set of protein−ligand complexes within a cluster an accurate superimposition of the proteins returns a meaningful overlay of the corresponding ligands. In this regard, there are many tools known that use the protein binding site, defined as a set of residues within a range of distance (5−6 Å) from the ligand atoms, to do the alignment. However, in cases where the binding site includes flexible regions, its mobility can impact the quality of the final overlay. If the residues selected to define the superimposition are present in different conformations in the different protein structures in the cluster, the result will be an average of the different conformations, and it can happen that ligand features binding to the same residues in different structures will be shifted apart from each other. The result of this is that when the ligands are viewed out of the context of the protein, pharmacophore features might be missed, for example in the adenosine deaminase (ADA) or coagulation factor VII discussed below.

We attempted to create a fully reproducible and automated alignment protocol, free from subjective interpretation over which residues to include. This is in contrast to other authors, who apply iterative superimposition algorithms, where the list of superimposed atoms pairs is continuously updated based on user defined criteria.[18,19] In addition, to avoid any bias from the protein flexibility, we propose an alternative protocol focused on protein domains or functional sites. More specifically, we looked for the existence of a PROSITE[24] motif located within or near the ligand binding site of the target under investigation. These motifs, typically around 10−20 amino acids in length, contain specific residues thought or proved to be important for the biological function of the protein that have been conserved in both structure and sequence during evolution. They arise because of particular requirements on the structure of specific regions of a protein that may be important, for example, for their binding properties or for their enzymatic activity. These requirements impose very tight constraints on the evolution of those relatively small but important portions of a protein sequence. PROSITE motifs are formulated like regular expressions, so that there is no confusion over which motifs will match to any particular PROSITE entry.

However, we included in our analysis some proteins even though the PROSITE motif is not available or is far from the binding site, which are interesting drug targets. In these cases, we manually selected a number of conserved residues based on the knowledge of the target. This follows the same principle as the use of a PROSITE[24] motif, but the grounds for using a particular set of residues is more subjective.

After identifying the residues to be used for the superimposition, we applied a fully automated procedure developed in-house. For each cluster, we needed to find a reference structure, so we ran an all-by-all comparison looped over all the

Table 2. Clusters of Targets Surviving the Filtering Process and Potentially Useful as Validation Sets[a]

ID[b]	target	no.	source	motif[c]	ref	RMSD (Å)
P00918	carbonic anhydrase II	135	*Homo sapiens*	PS00162	3qyk	0.079
P24941	cyclin-dependent kinase 2	105	*Homo sapiens*	VFEFL	1fvt	0.463
P00734	alpha thrombin	78	*Homo sapiens*	PS00135	3qwc	0.133
Q16539	p38 MAP kinase	75	*Homo sapiens*	VTHLM	3mvm	0.334
P07900	HSP 90-alpha	74	*Homo sapiens*	IVDTG	3b24	0.081
P00760	trypsin	71	*Bos taurus*	PS00135	1o33	0.116
P56817	beta-secretase 1	63	*Homo sapiens*	PS00141	4dju	0.966
O14757	serine/threonine-protein kinase Chk1	42	*Homo sapiens*	FLEYC	2x8d	0.346
P00489	protein (glycogen phosphorylase)	42	*Oryctolagus cuniculus*	PS00102	1xc7	0.108
P27487	dipeptidyl peptidase IV soluble form	39	*Homo sapiens*	PS00708	3d4l	0.196
P00742	coagulation factor XA	37	*Homo sapiens*	PS00135	2uwl	0.172
P11309	proto-oncogene serine/threonine-protein kinase Pim-1	31	*Homo sapiens*	ILERP	3c4e	0.105
P18031	protein (protein-tyrosine phosphatase 1b)	30	*Homo sapiens*	PS00383	2nta	0.122
P00749	protein (urokinase-type plasminogen activator)	27	*Homo sapiens*	PS00135	1gja	0.131
P03372	estrogen receptor	27	*Homo sapiens*	M-3-L-7-E-4-M-26-W-8-L-3-R-10-F	1xp9	0.155
P15121	aldose reductase	27	*Homo sapiens*	PS00798	3m4h	0.480
P00811	beta-lactamase	24	*Escherichia coli*	PS00336	1fcn	0.102
P09960	leukotriene A-4 hydrolase	23	*Homo sapiens*	PS00142	3fh5	0.067
Q13526	peptidyl-prolyl cis−trans isomerase NIMA-Interacting 1	23	*Homo sapiens*	PS01096	2xp5	0.146
P00517	cAMP-dependent protein kinase, alpha-catalytic subunit	21	*Bos taurus*	VMEY[VA]	2ojf	0.160
P0AE18	methionine aminopeptidase	21	*Escherichia coli*	PS00680	2evm	0.150
P28523	casein kinase II	19	*Zea mays*	IFEYV	2oxd	0.211
O15530	3-phosphoinositide dependent protein kinase-1	18	*Homo sapiens*	GLSYA	3nun	0.345
Q92731	estrogen receptor beta	18	*Homo sapiens*	M-3-L-7-E-4-M-26-W-8-L-3-R-10-F	1x76	0.177
P39900	macrophage metalloelastase	17	*Homo sapiens*	PS00142	3f18	0.050
P59071	phospholipase A2	17	*Daboia russellii pulchella*	PS00118	1th6	0.054
P10275	androgen receptor	16	*Homo sapiens*	L-7-E-4-M-34-L-3-R-10-F	2ax9	0.140
P14174	macrophage migration inhibitory factor	16	*Homo sapiens*	PS01158	3l5p	0.071
P53779	mitogen-activated protein kinase 10	16	*Homo sapiens*	VMELM	3ttj	0.317
P00374	dihydrofolate reductase	15	*Homo sapiens*	PS00075	1s3v	0.280
P14324	farnesyl pyrophosphate synthetase	15	*Homo sapiens*	PS00723	3n45	0.193
P78536	ADAM 17	15	*Homo sapiens*	PS00142	3ewj	0.080
P08581	hepatocyte growth factor receptor	14	*Homo sapiens*	VLPYM	3zxz	0.121
P25774	cathepsin S	14	*Homo sapiens*	PS00139	2r9n	0.077
P43235	cathepsin K	14	*Homo sapiens*	PS00139	1tu6	0.202
P68400	casein kinase II	14	*Homo sapiens*	VFEHV	3pe2	0.218
Q07343	cAMP-specific 3′,5′-cyclic phosphodiesterase 4B	14	*Homo sapiens*	PS00126	1xm6	0.103
Q08499	cAMP-specific 3′,5′-cyclic phosphodiesterase 4D	14	*Homo sapiens*	PS00126	1xor	0.064
Q9L5C8	beta-lactamase CTX-M-9	14	*Escherichia coli*	PS00146	3g32	0.053
O60674	tyrosine-protein kinase JAK2	13	*Homo sapiens*	IMEYL	3e62	0.093
P00523	proto-oncogene tyrosine-protein kinase Src	13	*Gallus gallus*	VTEYM	3f6x	0.106
P17612	cAMP-dependent protein kinase	13	*Homo sapiens*	VMEY[VA]	2gu8	0.164
P49841	glycogen synthase kinase-3 beta	13	*Homo sapiens*	VLDYV	3f7z	0.143
P11838	endothiapepsin	12	*Cryphonectria parasitica*	PS00141	3pi0	0.055
P24182	biotin carboxylase	12	*Escherichia coli*	PS00866	2w6z	0.318
P24627	lactotransferrin	12	*Bos taurus*	PS00207	3mjn	0.146
P45452	collagenase 3	12	*Homo sapiens*	PS00142	1xuc	0.068
Q9BJF5	calmodulin-domain protein kinase 1	12	*Toxoplasma gondii*	VGEVY	3i7c	0.122
O14965	serine/threonine-protein kinase 6	11	*Homo sapiens*	ILEYA	3h0y	0.120
P02829	HSP82	11	*Saccharomyces cerevisiae*	IRDSG	2xx2	0.052
P0A017	dihydrofolate reductase	11	*Staphylococcus aureus*	PS00075	3srs	0.123
P25440	bromodomain-containing protein 2	11	*Homo sapiens*	PS00633	2ydw	0.129
O60885	human BRD4	10	*Homo sapiens*	PS00633	3u5j	0.170
P00929	tryptophan synthase	10	*Salmonella typhimurium*	PS00167	1k3u	0.206
P06239	LCK kinase	10	*Homo sapiens*	ITEYM	3acj	0.311
P42330	aldo-keto reductase family 1 member C3	10	*Homo sapiens*	PS00798	3r8g	0.106
P47811	mitogen-activated protein kinase 14	10	*Mus musculus*	VTHLM	3p79	0.431
P51955	serine/threonine-protein kinase NEK2	10	*Homo sapiens*	VMEYC	2xkd	0.087
Q9QYJ6	phosphodiesterase-10A	10	*Rattus norvegicus*	PS00126	3hqz	0.096
O76074	cGMP-specific 3′,5′-cyclic phosphodiesterase	9	*Homo sapiens*	PS00126	3hdz	0.737

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

Table 2. continued

ID[b]	target	no.	source	motif[c]	ref	RMSD (Å)
O76290	pteridine reductase	9	*Trypanosoma brucei brucei*	PS00061	2x9n	0.092
P05326	isopenicillin n synthase	9	*Emericella nidulans*	PS00185	1qjf	0.077
P06401	progesterone receptor	9	*Homo sapiens*	L-3-L-7-Q-4-V-26-W-8-L-3-R-12-F	3zr7	0.159
P08254	stromelysin-1	9	*Homo sapiens*	PS00142	1hy7	0.093
P09955	procarboxypeptidase B	9	*Sus scrofa*	PS00132	2pj1	0.084
P28845	corticosteroid 11-beta-dehydrogenase isozyme 1	9	*Homo sapiens*	PS00061	3tfq	0.143
P50579	protein (methionine aminopeptidase)	9	*Homo sapiens*	PS01202	1yw7	0.105
P54760	ephrin type-B receptor 4	9	*Homo sapiens*	LTEFM	2vww	0.095
P56658	adenosine deaminase	9	*Bos taurus*	PS00485	1ndw	0.180
P61823	pancreatic ribonuclease A	9	*Bos taurus*	PS00127	3d6o	0.079
P80457	xanthine dehydrogenase	9	*Bos taurus*	PS00559	1vdv	0.094
Q581W1	pteridine reductase 1	9	*Trypanosoma brucei brucei*	PS00061	3jqa	0.099
Q9Y233	cAMP and cAMP-inhibited cGMP 3′, 5′-cyclic phosphodiesterase 10A	9	*Homo sapiens*	PS00126	3sni	0.191
P00520	proto-oncogene tyrosine-protein kinase ABL	8	*Mus musculus*	ITEFM	2qoh	0.127
P00730	carboxypeptidase A	8	*Bos taurus*	PS00132	1hdu	0.081
P00808	beta-lactamase	8	*Bacillus licheniformis*	PS00146	1i2w	0.111
P04058	acetylcholinesterase	8	*Torpedo californica*	PS00122	1gpk	0.078
P04642	L-lactate dehydrogenase A chain	8	*Rattus norvegicus*	PS00064	4ajl	0.051
P08069	insulin-like growth factor 1 receptor precursor	8	*Homo sapiens*	IMELM	3nw5	0.146
P0ABP9	purine nucleoside phosphorylase	8	*Escherichia coli*	PS01232	1pr2	0.073
P0C5C1	beta-lactamase	8	*Mycobacterium tuberculosis*	PS00146	3n8l	0.062
P15090	fatty acid-binding protein, adipocyte	8	*Homo sapiens*	PS00214	3p6h	0.129
P22906	dihydrofolate reductase	8	*Candida albicans*	PS00075	1ia1	0.151
P35968	vascular endothelial growth factor receptor 2	8	*Homo sapiens*	IVEFC	1y6a	0.136
P41148	endoplasmin	8	*Canis lupus familiaris*	VTDTG	2gqp	0.059
P42574	caspase-3	8	*Homo sapiens*	PS01121	2xyg	0.084
Q00511	uricase	8	*Aspergillus flavus*	PS00366	1wrr	0.095
Q02127	dihydroorotate dehydrogenase, mitochondrial	8	*Homo sapiens*	PS00911	3kvj	0.139
Q9T0N8	cytokinin dehydrogenase 1	8	*Zea mays*	PS00862	3bw7	0.211
P04035	protein (HMG-COA reductase)	7	*Homo sapiens*	PS00066	1hw9	0.085
P07688	cathepsin B	7	*Bos taurus*	PS00139	2dc9	0.045
P08235	mineralocorticoid receptor	7	*Homo sapiens*	L-3-L-7-Q-4-V-26-W-8-L-3-R-12-F	2aa6	0.118
P16184	dihydrofolate reductase	7	*Pneumocystis carinii*	PS00075	3nz9	0.328
P25779	cruzain	7	*Trypanosoma cruzi*	PS00139	1me3	0.136
P28482	mitogen-activated protein kinase 1	7	*Homo sapiens*	VQDLM	3i60	0.274
P30291	wee1-like protein kinase	7	*Homo sapiens*	QNEYC	2in6	0.081
P30405	peptidyl-prolyl cis−trans isomerase F, mitochondrial	7	*Homo sapiens*	PS00170	3rdb	0.065
P35557	glucokinase isoform 2	7	*Homo sapiens*	PS00378	3vev	0.126
P52700	metallo-beta-lactamase L1	7	*Stenotrophomonas maltophilia*	PS00743	2fu9	0.065
Q57834	tyrosyl-tRNA synthetase	7	*Methanocaldococcus jannaschii*	PS00178	2hgz	0.147
A9JQL9	dehydrosqualene synthase	6	*Staphylococcus aureus*	PS01044	3acy	0.099
P00509	aspartate aminotransferase	6	*Escherichia coli*	PS00105	1cq7	0.063
P00797	renin	6	*Homo sapiens*	PS00141	2g1r	0.103
P09467	fructose-1,6-bisphosphatase 1	6	*Homo sapiens*	PS00124	3kc1	0.105
P0A5J2	methionine aminopeptidase	6	*Mycobacterium tuberculosis*	PS00680	3pka	0.100
P0AD64	beta-lactamase SHV-1	6	*Klebsiella pneumoniae*	PS00146	2a49	0.072
P11509	cytochrome P450, family 2, subfamily A, polypeptide 6	6	*Homo sapiens*	PS00086	1z10	0.076
P51857	3-oxo-5-beta-steroid 4-dehydrogenase	6	*Homo sapiens*	PS00798	3g1r	0.068
Q10714	angiotensin converting enzyme	6	*Drosophila melanogaster*	PS00142	2x93	0.065
P00469	thymidylate synthase	5	*Lactobacillus casei*	PS00091	3ik1	0.151
P00772	elastase	5	*Sus scrofa*	PS00135	1e34	0.080
P08709	coagulation factor VII	5	*Homo sapiens*	PS00135	2ec9	0.337
P12758	uridine phosphorylase	5	*Escherichia coli*	PS01232	1u1d	0.064
P23470	receptor-type tyrosine-protein phosphatase gamma	5	*Homo sapiens*	PS00383	3qck	0.071
P36897	TGF-beta receptor type I	5	*Homo sapiens*	VSDYH	3hmm	0.087
P48736	phosphatidylinositol-4,5-bisphosphate 3-kinase	5	*Homo sapiens*	PS00915	3l54	0.293
Q04771	activin receptor type-1	5	*Homo sapiens*	ITHYH	3q4u	0.085
Q3JRA0	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	5	*Burkholderia pseudomallei*	PS01350	3k14	0.049

Table 2. continued

ID[b]	target	no.	source	motif[c]	ref	RMSD (Å)
Q9BZP6	acidic mammalian chitinase	5	*Homo sapiens*	PS01095	3rme	0.058

[a]The UniProt ID, target's name, number of remaining structures, motif used for the superimposition, and average RMSD of the reference structure are indicated. [b]UniProt ID. [c]PROSITE motif or a sequence of residues (even not consecutive) selected on the basis of the knowledge of the target.

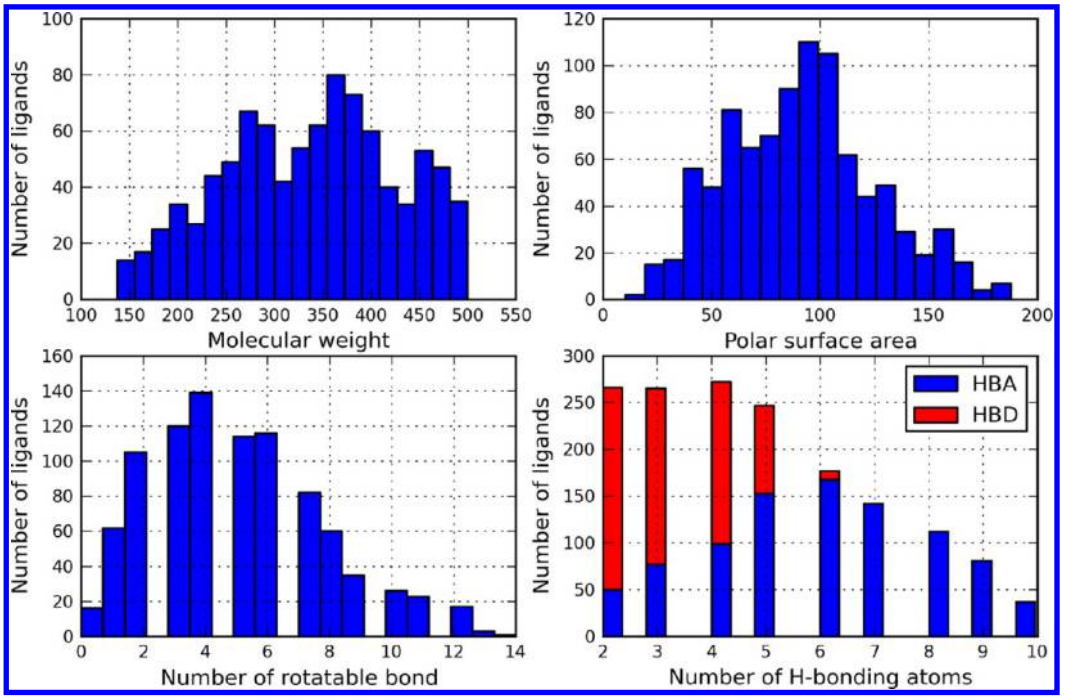


Figure 1. Distributions of physicochemical properties of the ligands in the pharmacophore validation set.

sequences, and we calculated the minimized RMSD value after superposition. For pairs of proteins with multiple chains of identical sequence in the structure, we considered the combination of chains that gave the lowest RMSD. The chain with the lowest mean RMSD to the other structures was used as the reference structure, and the final overlay produced by superimposition onto this reference chain. In all cases, the protein structures were superimposed using the backbone atom coordinates of the selected residues. All overlays were performed using the OERMSD function from OEChem.[25]

Setting Up Complexes. Even working with high-resolution X-ray structures, it is worth fixing common problems such as missing hydrogen atoms, ambiguous protonation states, and flipped residues as an initial step in any drug design project. In this work, we prepared each PDB structure with the Protein Preparation Wizard implemented in Maestro 9.2.[26]

In the first stage of preprocessing a structure, the program assigned bonds and bond orders, added hydrogens, replaced bonds to metals with zero-order bonds, adjusted the formal charge on the metal and the neighboring atoms, converted selenomethionines (MSE) to methionines (MET), and deleted waters that were more than 5 Å from any HET group. We used Epik[27] to predict ionization and tautomeric states of the ligands within the protein binding site in the pH range specified by default at 7 ± 4 units. Finally, we employed the automated optimization of the hydrogen-bonding network that reorients hydroxyl and thiol groups, water molecules, amide groups of asparagine and glutamine, and the imidazole ring in histidine; predicts the protonation states of histidine, aspartate and glutamate; and assigns the tautomeric states of histidine. All other protein and ligand atoms were frozen in position.

Subsetting Clusters. It is difficult to derive robust pharmacophore models when dealing with large data sets, and most pharmacophore elucidation algorithms cannot cope with a large number of input structures. In cases where we found more than 40 ligands in complex with the same protein, we decided to use a systematic approach to reduce the quantity of data. We did this on the assumption that pharmacophore elucidation is a combinatorial problem, and hence, large data sets provide a formidable barrier to validation. This is certainly borne out by publications to date, which generally employ quite small validation sets. Only the recent FLAPharm paper has defined overlay sets of significant size. It would be unfortunate if we took care to define an extensive validation set, with supporting benchmarks, that was then reduced in size by others due to algorithmic constraints, or even worse, was ignored as being beyond the capability of a novel but effective algorithm.

For each protein−ligand complex within the large clusters (i.e., more than 40 ligands), we performed a contact analysis to detect key residues for the recognition of a specific binding site. We used an in-house algorithm written in MATLAB[28] to measure all atom−atom distances between protein and ligand. Proton positions were not assigned. A ligand atom was defined as being in contact with a protein atom if their separation was 3.8 Å or less. This value was chosen because it represents the S−S contact distance according to the Bondi scale and should therefore capture most contacts in protein−ligand systems.[29] Metals and waters were considered to be part of the protein if they contacted both protein and ligand (with no requirement

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

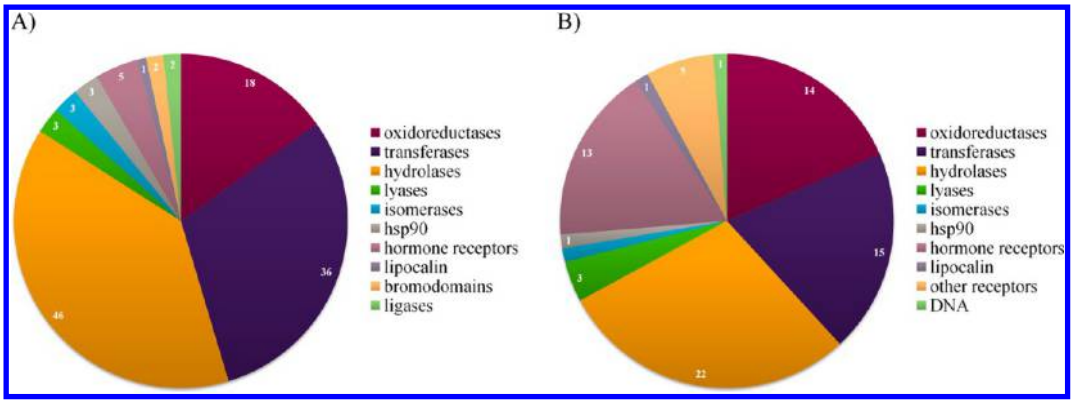


Figure 2. Protein classes represented in the current data set (A) compared with those in PharmBench (B).

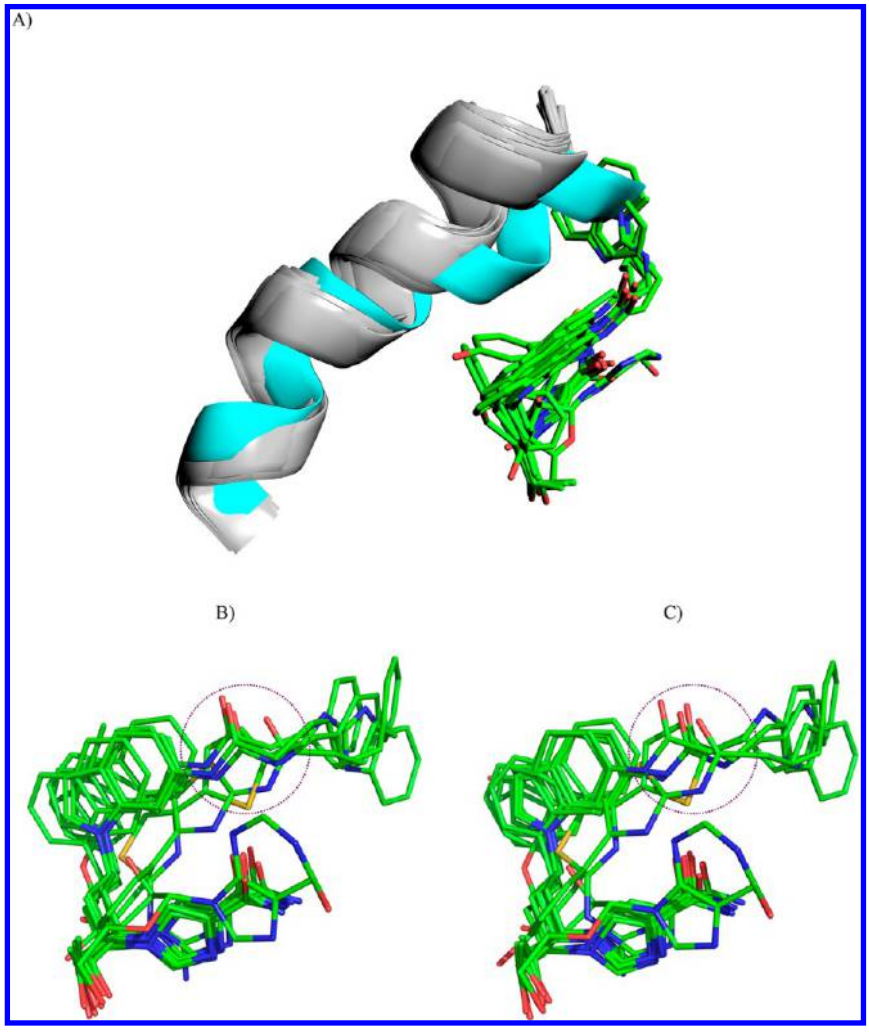


Figure 3. (A) α-Helix corresponding to the structural gate of the adenosine deaminase (ADA) active site. The PDB entry 1krm shows a closed form of the helix (cyan ribbon), while the other ADA protein–ligand complexes are all in an open conformation (white ribbon). The ligand overlay is rendered as green sticks. Overlay obtained using the PROSITE motif (B) compared with that obtained through relibase+ (C). The first shows that some features of the same type are better aligned.

for a direct protein–ligand contact in that case). This approach is very simple and reproducible, requiring no preprocessing of the PDB data; it does not require any further breakdown into detailed contact types, such as hydrogen bonds or hydrophobic contacts. Once we had identified and categorized contacts by residue, we were able to tabulate the results and identify the most contacted residues and the percentage of ligands contributing to a given contact. We then followed a systematic procedure for subsetting clusters: starting from the most contacted residue for that ligand set, we excluded ligands that did not interact with this residue; then we updated the contact statistics based on the reduced list of complexes and iterated. This sequence stopped when the list of ligands was less than 40, and the next residue was contacted by less than 75% of ligands.

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852–866

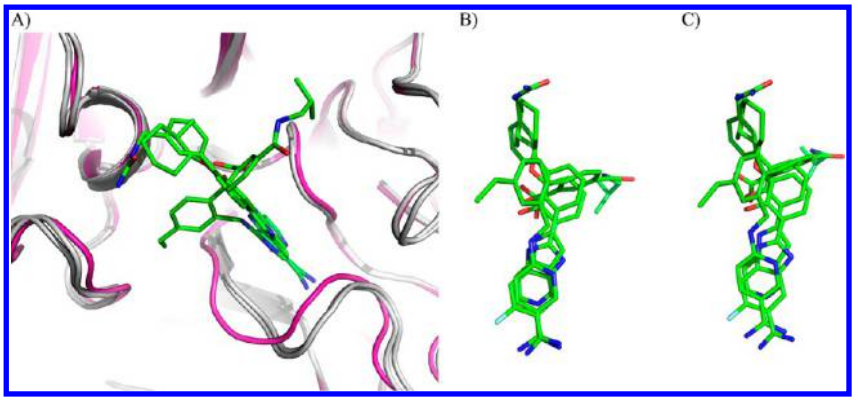


Figure 4. (A) Binding site of coagulation factor VII. The flexible loop highlighted in the figure adopts a different conformation when comparing the PDB entry 2flr (magenta ribbon) with the other two structures (white ribbon). The ligand overlay is rendered as green sticks. Overlay obtained using the PROSITE motif (B) compared with that obtained using the CE algorithm as implemented in PyMOL (C).

Overlay Analysis. Three parameters were used to score the final overlay sets. A Python script using the OEShape toolkit was used to calculate molecular shape[30] and the color score for all the $n(n-1)/2$ pairwise ligand combinations within a set by disabling the solid-body optimization process to maintain unchanged the starting conformation. We thus measured the average values of shape and color similarities. They both range between 0 and 1, and the higher the similarity value the more similar the two molecules according to that descriptor.

In addition, the Tanimoto coefficient was used to measure the 2D fingerprint similarity, using our in-house foyfi fingerprint.[31] Again, the average value was calculated on the basis of an all-by-all comparison of ligands in each overlay.

The aim of this analysis was to assess the quality of an overlay for pharmacophore validation purposes, the idea being that those overlays with a good shape or color score should be easier for a pharmacophore generation program to reproduce. If the overlay defined by the protein structures has no single feature in common with at least some of the other ligands, it is unrealistic to expect a pharmacophore program to find the correct solution. To this end, consensus-based methods were investigated with the aim of distinguishing good and poor overlays. Accordingly, different ranking rules (shape and color similarities sorted in a descending order and the 2D fingerprint similarity sorted in ascending order) were applied to prioritize molecular overlays. The maximum rank of the three was used for each set as the global rank, meaning that the higher the global rank the poorer the overlay as a validation set. The maximum rank was used for the consensus score, rather than, for example, the sum of ranks, so as to penalize overlay sets where the ligands are all very similar. Such sets have very good scores for color score and shape matching but are still of limited utility in validating overlay programs.

RESULTS AND DISCUSSION

Validation Set. Table 2 shows the 119 targets identified after the final filtering, annotated with a PROSITE[24] motif placed near the binding site. All the structures within each of these clusters were automatically superimposed onto a selected reference chain by using the backbone atom coordinates of residues corresponding to the specified motif (see Materials and Methods). Every protein−ligand complex was optimized through the Protein Preparation Wizard tool from Maestro 9.2[26] in order to adjust the protonation state of the ligand in protein context. Lastly, the ligand coordinates from each superimposed complex were extracted to obtain the final alignment which will be referred to as the target overlay.

At this stage, to guarantee a high quality set of structures, we carefully looked at each single ligand. We fixed those cases where the automated protocol failed in assigning bond orders or protonation state, and we flagged up any problems of strained geometry or questionable conformation in Table S2 of the Supporting Information. In the latter case, we decided to keep these structures because it will not impact the feature assignment.

The visual inspection of the resulting overlays allowed us also to identify a few cases where ligands are bound in an allosteric pocket. In the latter situation, if more than five ligands were overlaid in the additional site, this set was considered as an additional overlay, otherwise these ligands were excluded from the analysis. Examples of targets complexed with allosteric inhibitors are farnesyl pyrophosphate synthetase (UniProt ID: P14324) and protein kinase A (UniProt ID: O15530).

The distribution of key physicochemical properties in the data set is given in Figure 1.

Although a similar benchmarking set has been published recently, the criteria used to select and filter targets allowed us to compile a more comprehensive data set, not only in terms of number of targets and ligands but also in terms of protein families represented and ligand diversity explored. Pie charts in Figure 2 show the distribution of protein classes represented in the current data set and in PharmBench. Hydrolase, transferases, and oxidoreductases are the top three most populated families in both data sets. Notably, our data set contains ligases (i.e., UniProt IDs: P24182 and Q57834) and bromodomains (i.e., UniProt IDs: O60885 and P25440) that are absent from PharmBench. The latter are epigenetic proteins intensively pursued as targets in drug discovery.[32] The filtering criterion based on the availability of PROSITE motifs led us to the exclusion of a number of receptors that are present in PharmBench, although we have included five hormone receptors.

As an additional analysis, we have compared the average fingerprint similarity (calculated as detailed above) of ligands in the 33 common targets. All data are stored in Table S3 of the Supporting Information. The most diverse sets of ligands in our data set, if compared with the corresponding sets in PharmBench, are mineralocorticoid receptor (UniProt ID: P08235), estrogen receptor beta (UniProt ID: Q92731), and mitogen-activated protein kinase 1 (UniProt ID: P28482). The

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

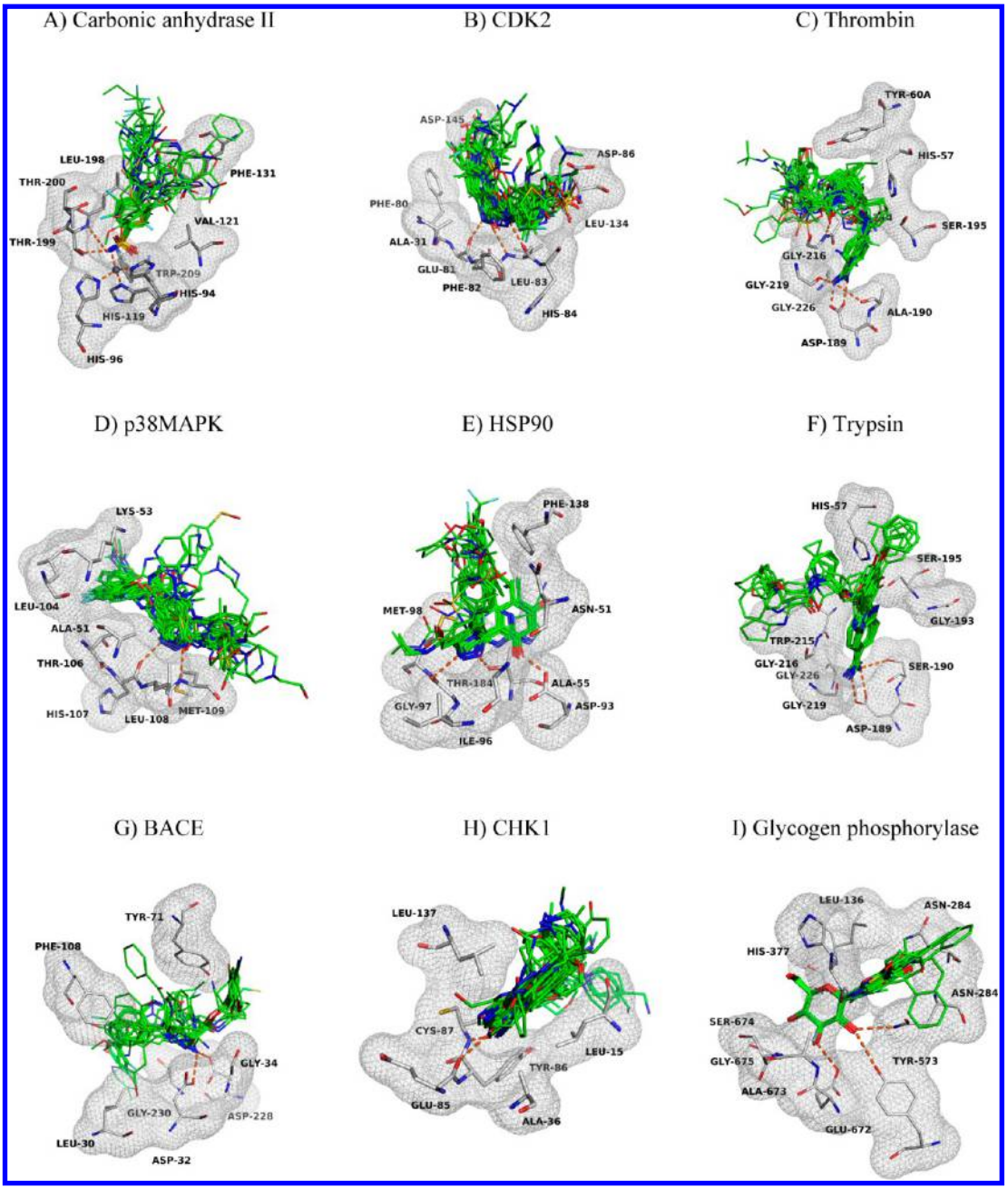


Figure 5. Final overlays of clusters with more than 40 structures selected by contact analysis. Residues contacted by all ligands in each data set are shown. Key hydrogen bonds are represented as orange dashed lines.

difference values of the average 2D fingerprint similarity are 14.54%, 13.05%, and 11.08%, respectively. The only target that appears to be more diverse in PharmBench is renin (UniProt ID: P00797) at 15.80%. However, the big contribution to this diversity is due to ligand 3k1w that is absent from our set because it failed the Lipinski rules cut off. It is worth noting that the scoring procedure discussed below ranked this target among the poorest sets for pharmacophore validation, while estrogen receptor beta, which is the second most diverse set, is also the second best ranked overlay according to the same scoring.

Overlaying onto a PROSITE Motif. Unlike previously published work, we used a set of residues corresponding to the PROSITE[24] motif of a given target to do the superimposition of protein structures in order to avoid biases from protein flexibility. Adenosine deaminase (UniProt ID: P56658) is a good example to show how different conformations of the binding site influence the final overlay.

As shown in Table 2, our set contains nine ligands (PDB codes: 1krm, 1ndz, 1ndw, 1ndy, 1o5r, 1uml, 1v7a, 1v79, 2e1w) which are all in common with the corresponding set from Taylor et al.[17] bar one, 1ndz. Three ligands from that data set did not pass our filtering process because they contain four methylene linkers (1qxl and 1wxy) or fail Lipinski's rule (1ndv). However, to make things consistent, we ran our protocol on the same set of 11 protein−ligand complexes belonging to Taylor's data set of ADA and did the superimposition onto the residues included in the motif PS00485. The reference structure was now 1qxl with a mean RMSD value equal to 0.185. Figure 3 shows the comparison

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

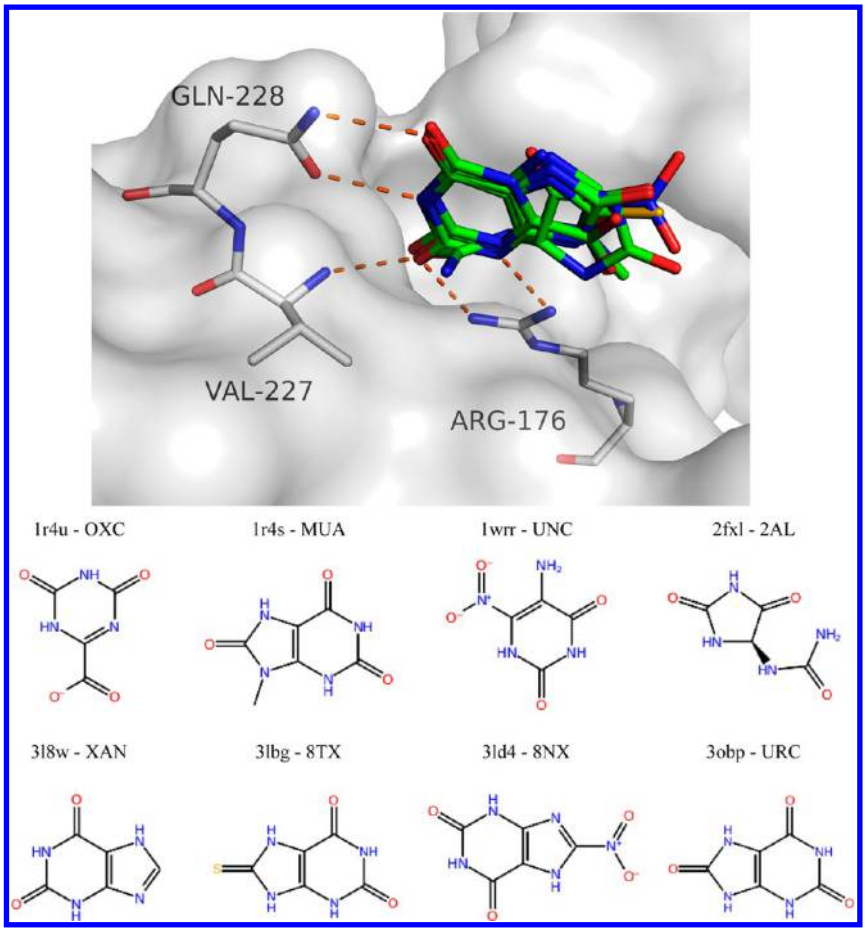


Figure 6. Molecular overlay of ligands complexed with the protein drug uricase. Conserved interactions with the binding site are shown as orange dashed lines, while the protein is rendered as a surface. Cartesian coordinates are taken from the reference structure (PDB code 1wrr). Two-dimensional representation of each ligand is also shown, PDB code and ligand ID are indicated.

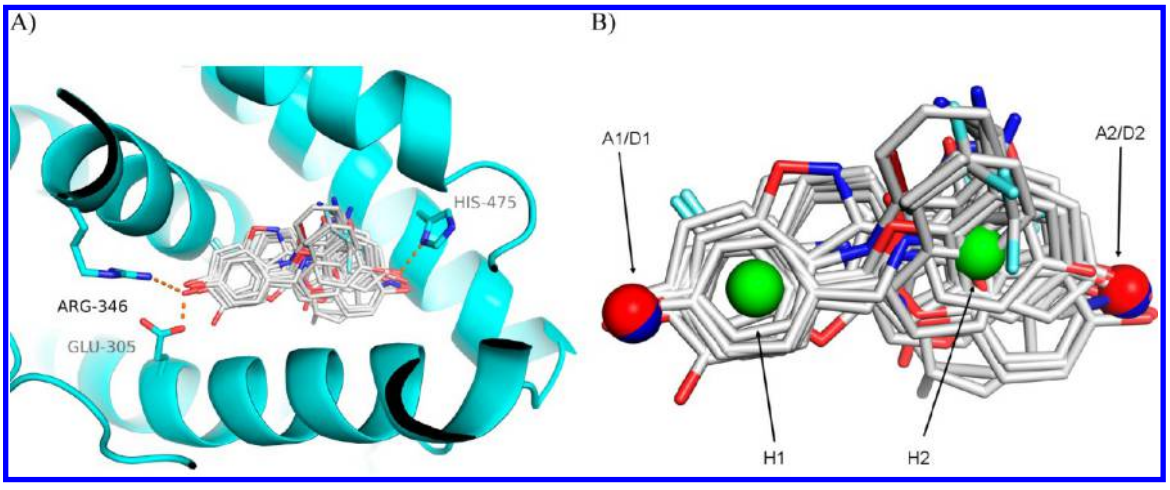


Figure 7. (A) Molecular overlay of 18 ligands of the estrogen receptor beta superimposed onto the 3D-coordinates of 1 × 76. Important residues are rendered as stick. (B) Four pharmacophoric points are represented as spheres. A1, D1, H1, and H2 are full pharmacophoric points, while A2 and D2 are partial. Hydrophobic, donor, and acceptor groups are colored in green, blue, and red, respectively.

between our overlay (B) and the overlay published by Taylor et al. (C) that was obtained by least-squares fitting of the binding site atoms through relibase+.[33] ADA has two distinct conformations, named the open and the closed forms, due to an inhibitor induced-fit phenomenon. The 11 structures under investigation are all in an open form except for 1krm, whose α-helix, consisting of the structural gate of the active site, adopts a different conformation (Figure 3A).[34] Therefore, maximizing the superimposition of the whole binding site, including this flexible moiety, returns a final overlay where similar chemical groups from different ligands are not well overlaid (e.g., the amido group from 1uml, 1qxl, and 1wxy). In contrast, our protocol yielded a better match of common features.

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

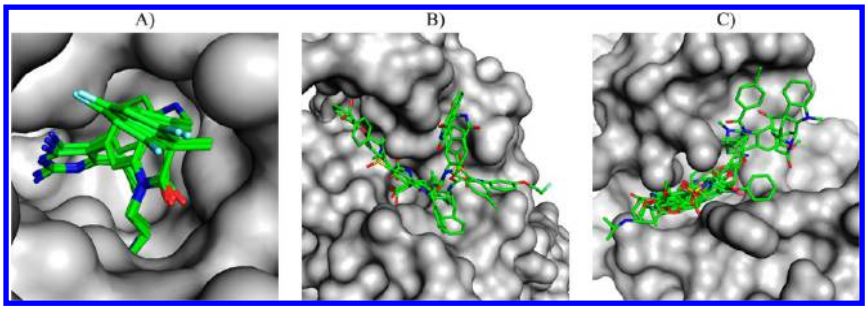


Figure 8. Molecular overlays which we deem the poorest validation sets. (A) Five ligands complexed with renin are very well overlaid but show high 2D-fingerprint similarity. Seven ligands complexed with caspase 3 (B) and 16 ligands complexed with phospholipase A2 (C) all bind in the same region of the protein but show low shape matching and low feature similarity score.

Another example is coagulation factor VII (UniProt ID: P08709) for which we provide a set of five structures overlaid onto the residues included in the motif PS00135. Three of these structures are common to the corresponding set from PharmBench (PDB codes: 1o5d, 2ec9, 2flr). A visual inspection revealed that a flexible loop is embedded in the ligand binding site (Figure 4A), and specifically, the PDB entry 2flr adopts a different conformation of this loop if compared with the other two structures. Figure 4 shows the comparison between our overlay (B) and the overlay included in PharmBench (C). Again, our approach resulted in a better match of common features, especially the benzamidine core, which is known to interact with Asp189 (part of PS00135) at the bottom of the S1 pocket of trypsin-like serine proteases.

Subsetting Highly Populated Clusters. The contact-based protocol described above was used to subset nine targets complexed with more than 40 different ligands able to pass the filtering process.

Carbonic Anhydrase II. The original set made of 135 ligands was reduced to 24 that are all able to contact the Zn^{2+} ion, Gln92, His94, His96, His119, Val121, Phe131, Leu198, Thr199, Thr200, and Trp209 (Figure 5A). All the ligands in this data set bind in essentially the same way, with a sulfonamide group coordinating a zinc ion and forming two key hydrogen bonds with the Thr199. The other selected residues are involved in polar or hydrophobic interactions but to some extent define the shape of the pocket frequently explored to target that protein.

Taylor et al.[17] used a set of 13 ligands to validate their algorithm. Among them, two structures (PDB codes: 1oq5 and 1zh9) are also in our data set. Carbonic anhydrase is also present in PharmBench with 84 structures, but 12 ligands from our filtered set of 24 are not in common with it.

CDK2. Starting from 105 ligands, we went down to 24 after analyzing contacts in the binding site. We kept ligands if they contact Ile10, Ala31, Phe80, Glu81, Phe82, Leu83, His84, Asp86, Leu134, and Asp145 (Figure 5B). Key interactions were the two hydrogen bonds with the backbone of Glu81 and Leu83, both belonging to the hinge moiety.

CDK2 has been used in the past as validation data set for pharmacophore elucidation. Patel et al.[7] presented a set of six ligands, while Jones[15] used two sets of 10 (diverse set) and nine (focused set) ligands to validate GAPE. Our data set has only one structure (PDB code: 1di8) in common with the Patel data set and two structures (PDB codes: 1di8 and 2bhe) with the Jones set.[15] PharmBench includes the same ligands as in Patel et al.

Thrombin. We selected 28 ligands from the 78 that passed the filtering process. They all contact His57, Tyr60, Asp189, Ala190, Ser195, Ser214, Trp215, Gly216, Gly219, and Gly226 (Figure 5C). The ionic interaction with the carboxylic group of Asp189, together with the hydrogen bonds to the backbone of Gly216 and Gly219, seemed to be conserved.

Thrombin is another target included in the Patel data set,[7] but none of those seven ligands are present in our data set. This set of seven ligands is also included in PharmBench in addition to another set of 41 structures. Only 10 ligands from our reduced set of 28 are also represented among the 41.

p38MAPK. Only 27 ligands out of 75 are in the final set for p38MAPK. As expected, these ligands share the hydrogen bonds with the backbone of His107 and Met109, part of the hinge region. Additional residues contacted are Ala51, Lys53, Leu104, and Leu108 (Figure 5D).

HSP90. The contact analysis permitted us to select 24 ligands from the original set of 74. Important interactions shared by most of them are the hydrogen bonds with Asp93, Gly97, and Thr184. However, we also included ligands able to contact Asn51, Ala55, Ile96, Met98, and Phe138 via polar or hydrophobic interactions (Figure 5E).

Three of the ligands (PDB codes: 1yc1, 2bsm, and 2cct) are included in the 10 from the Taylor's[17] data set. Eleven ligands are not included in the 57 from PharmBench.

Trypsin. Starting from 71 structures, we drew a data set of 22 ligands, all able to contact His57, Asp189, Ser190, Gln192, Ser195, Trp215, Gly216, Gly219, and Gly226 (Figure 5F). It is known for this serine protease that inhibitors are anchored in the S1 pocket via a network of hydrogen bonds to side chains of Asp189, Ser190, and the backbone of Gly219. In addition, the hydrogen bond to Ser195 also appears to be a conserved interaction. Trypsin is another target in common with PharmBench which contains 83 structures, but 10 ligands from our reduced set are absent from it.

BACE. We selected 18 ligands from 63, which all interact with Leu30, Asp32, Gly34, Tyr71, Phe108, Asp228, and Gly230. As expected, hydrogen bonds to the two aspartate residues (Asp32 and Asp228) in the active site are conserved across multiple ligands as well as those with the backbone of Gly34 and Gly230 (Figure 5G).

CHK1. We excluded 35 ligands from the original pool of 42 to define a set of 14 ligands able to contact Leu15, Ala36, Glu85, Tyr86, Cys87, and Leu137 (Figure 5H). As for the other two protein kinases discussed above, the hydrogen bonds to the backbone of Glu85 and Cys87 from the hinge region are crucial to target this enzyme.

Glycogen Phosphorylase. We made a definitive set of 20 ligands from 42, all contacting Gly135, Leu136, Asn274, Asn274, His377, Tyr573, Glu672, Alal673, Ser674, and Gly675

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

Table 3. Final Overlays (121) Sorted by the Consensus Maximum Calculated over the Ranking Position of the Average Color Score, Average Shape Similarity, and Average 2D Fingerprint Similarity[a]

ID[b]	target	no.	average 2D fingerprint similarity	average shape similarity	average color score
Q00511	uricase	8	0.412	0.845	0.446
Q92731	estrogen receptor beta	18	0.433	0.720	0.281
P30405	peptidyl-prolyl cis−trans isomerase F, mitochondrial	5	0.354	0.603	0.278
P14324	farnesyl pyrophosphate synthetase	8	0.455	0.718	0.434
P04035	protein (HMG-COA reductase)	7	0.450	0.534	0.398
P10275	androgen receptor	11	0.446	0.610	0.210
P15090	fatty acid-binding protein, adipocyte	8	0.440	0.551	0.202
P03372	estrogen receptor	27	0.467	0.632	0.222
Q581W1	pteridine reductase 1	8	0.478	0.608	0.415
P11509	cytochrome P450, family 2, subfamily A, polypeptide 6	6	0.293	0.609	0.186
P00929	tryptophan synthase	10	0.491	0.594	0.271
P0AE18	methionine aminopeptidase	21	0.398	0.498	0.168
P00918	carbonic anhydrase II	23	0.444	0.471	0.278
P25774	cathepsin S	14	0.496	0.490	0.179
P00749	protein (urokinase-type plasminogen activator)	27	0.461	0.471	0.159
Q9T0N8	cytokinin dehydrogenase 1	8	0.497	0.695	0.220
P41148	endoplasmin	8	0.492	0.464	0.197
P51857	3-oxo-5-beta-steroid 4-dehydrogenase	6	0.463	0.667	0.158
P07900	HSP 90-alpha	23	0.464	0.461	0.183
Q13526	peptidyl-prolyl cis−trans isomerase NIMA-Interacting 1	23	0.455	0.461	0.155
P09955	procarboxypeptidase B	9	0.503	0.523	0.346
P25440	bromodomain-containing protein 2	11	0.391	0.518	0.151
P24182	biotin carboxylase	12	0.403	0.455	0.200
P24941	cyclin-dependent kinase 2	24	0.496	0.533	0.150
O60885	human BRD4	10	0.506	0.595	0.218
P0C5C1	beta-lactamase	8	0.509	0.454	0.269
P08581	hepatocyte growth factor receptor	13	0.491	0.449	0.160
P45452	collagenase 3	12	0.509	0.449	0.157
P18031	protein (protein-tyrosine phosphatase 1b)	30	0.415	0.447	0.141
Q02127	dihydroorotate dehydrogenase, mitochondrial	8	0.516	0.675	0.167
P43235	cathepsin K	13	0.482	0.439	0.167
P08235	mineralocorticoid receptor	7	0.527	0.819	0.405
P78536	ADAM 17	15	0.504	0.435	0.135
P35557	glucokinase isoform 2	7	0.529	0.470	0.137
P39900	macrophage metalloelastase	17	0.531	0.538	0.216
P09960	leukotriene A-4 hydrolase	19	0.477	0.610	0.133
O60674	tyrosine-protein kinase JAK2	13	0.473	0.434	0.145
O14965	serine/threonine-protein kinase 6	11	0.533	0.444	0.190
O14757	serine/threonine-protein kinase Chk1	14	0.489	0.465	0.133
P02829	HSP82	11	0.533	0.509	0.325
P15121	aldose reductase	27	0.436	0.430	0.162
Q9L5C8	beta-lactamase CTX-M-9	14	0.416	0.425	0.235
P0A5J2	methionine aminopeptidase	6	0.442	0.425	0.213
P25779	cruzain	7	0.486	0.480	0.126
P08254	stromelysin-1	9	0.556	0.496	0.235
P06401	progesterone receptor	9	0.399	0.559	0.122
P00517	cAMP-dependent protein kinase, alpha-catalytic subunit	21	0.522	0.411	0.145
P09467	fructose-1,6-bisphosphatase 1	6	0.556	0.455	0.227
Q16539	p38 MAP kinase	27	0.502	0.486	0.121
Q9QYJ6	phosphodiesterase-10A	10	0.544	0.403	0.136
O15530	3-phosphoinositide dependent protein kinase-1	14	0.559	0.497	0.183
Q07343	cAMP-specific 3′,5′-cyclic phosphodiesterase 4B	14	0.484	0.498	0.116
P00509	aspartate aminotransferase	6	0.561	0.633	0.332
P56817	beta-secretase 1	18	0.525	0.402	0.174
P28523	casein kinase II	19	0.358	0.456	0.114
P04642	L-lactate dehydrogenase A chain	8	0.515	0.400	0.214
Q08499	cAMP-specific 3′,5′-cyclic phosphodiesterase 4D	14	0.442	0.435	0.112
Q04771	activin receptor type-1	5	0.579	0.551	0.255
P35968	vascular endothelial growth factor receptor 2	8	0.559	0.396	0.151
P00760	trypsin	22	0.587	0.507	0.317
P27487	dipeptidyl peptidase IV soluble form	39	0.495	0.395	0.128
P00730	carboxypeptidase A	8	0.588	0.626	0.193

dx.doi.org/10.1021/ci400020a | J. Chem. Inf. Model. 2013, 53, 852−866

Table 3. continued

ID[b]	target	no.	average 2D fingerprint similarity	average shape similarity	average color score
P36897	TGF-beta receptor type I	5	0.590	0.694	0.268
P00742	coagulation factor XA	37	0.596	0.622	0.141
P80457	xanthine dehydrogenase	9	0.330	0.469	0.100
P00734	alpha thrombin	28	0.602	0.580	0.199
P52700	metallo-beta-lactamase L1	6	0.337	0.383	0.118
P00489	protein (glycogen phosphorylase)	22	0.604	0.792	0.589
P17612	cAMP-dependent protein kinase	13	0.480	0.381	0.138
P06239	LCK kinase	10	0.458	0.411	0.098
P22906	dihydrofolate reductase	8	0.608	0.553	0.560
P49841	glycogen synthase kinase-3 beta	13	0.433	0.380	0.129
P00520	proto-oncogene tyrosine-protein kinase ABL	5	0.565	0.407	0.097
P14324*	farnesyl pyrophosphate synthetase	7	0.447	0.387	0.089
P08069	insulin-like growth factor 1 receptor precursor	8	0.632	0.378	0.173
P08709	coagulation factor VII	5	0.537	0.355	0.153
P68400	casein kinase II	14	0.357	0.428	0.087
P16184	dihydrofolate reductase	7	0.650	0.594	0.367
P00523	proto-oncogene tyrosine-protein kinase Src	12	0.556	0.349	0.106
Q9BZP6	acidic mammalian chitinase	5	0.390	0.411	0.086
Q9BJF5	calmodulin-domain protein kinase 1	12	0.650	0.680	0.396
Q9Y233	cAMP and cAMP-inhibited cGMP 3′, 5′-cyclic phosphodiesterase 10A	9	0.503	0.443	0.085
Q10714	angiotensin converting enzyme	6	0.653	0.657	0.438
P47811	mitogen-activated protein kinase 14	10	0.526	0.347	0.111
P50579	protein (methionine aminopeptidase)	9	0.386	0.393	0.080
P51955	serine/threonine-protein kinase NEK2	10	0.654	0.734	0.459
P42330	aldo-keto reductase family 1 member C3	10	0.388	0.337	0.100
P48736	phosphatidylinositol-4,5-bisphosphate 3-kinase	5	0.410	0.445	0.078
P0A017	dihydrofolate reductase	11	0.657	0.648	0.465
O76290	Pteridine reductase	9	0.423	0.318	0.273
O15530*	3-phosphoinositide dependent protein kinase-2	5	0.433	0.439	0.075
P00374	dihydrofolate reductase	15	0.659	0.598	0.369
P0AD64	beta-lactamase SHV-1	6	0.497	0.317	0.141
P23470	receptor-type tyrosine-protein phosphatase gamma	5	0.680	0.775	0.347
P11309	proto-oncogene serine/threonine-protein kinase Pim-1	31	0.372	0.463	0.074
P12758	uridine phosphorylase	5	0.698	0.661	0.492
P00808	beta-lactamase	8	0.404	0.307	0.107
P53779	mitogen-activated protein kinase 10	16	0.501	0.387	0.073
P30291	wee1-like protein kinase	7	0.700	0.784	0.495
P04058	acetylcholinesterase	8	0.408	0.306	0.062
P54760	ephrin type-B receptor 4	9	0.702	0.698	0.443
P28482	mitogen-activated protein kinase 1	7	0.537	0.300	0.106
P28845	corticosteroid 11-beta-dehydrogenase isozyme 1	9	0.465	0.479	0.061
P00469	thymidylate synthase	5	0.704	0.485	0.127
O76074	cGMP-specific 3′,5′-cyclic phosphodiesterase	9	0.619	0.348	0.058
P05326	isopenicillin n synthase	9	0.731	0.778	0.540
P14174	macrophage migration inhibitory factor	16	0.370	0.310	0.054
Q57834	Tyrosyl-tRNA synthetase	7	0.757	0.682	0.377
Q3JRA0	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4	0.367	0.280	0.124
P00811	beta-lactamase	22	0.387	0.290	0.053
P56658	adenosine deaminase	9	0.759	0.548	0.368
P00772	elastase	5	0.455	0.278	0.134
A9JQL9	dehydrosqualene synthase	6	0.349	0.280	0.029
P0ABP9	purine nucleoside phosphorylase	8	0.763	0.861	0.414
P11838	endothiapepsin	11	0.352	0.246	0.074
P07688	cathepsin B	7	0.795	0.640	0.385
P24627	lactotransferrin	12	0.351	0.178	0.016
P59071	phospholipase A2	16	0.343	0.144	0.015
P00797	renin	5	0.841	0.715	0.509
P61823	pancreatic ribonuclease A	7	0.856	0.542	0.275
P42574	caspase-3	7	0.440	0.130	0.020

[a]Two targets have an additional overlay of ligands bound in an allosteric site and are labelled with a star. [b]UniProt ID.

(Figure 5I). A key interaction was the hydrogen bond to the hydroxyl group of Tyr573, but the hydrogen bonds to Asn284, Glu672, and Gly675 are also fairly well conserved.

Good or Poor Overlays for Pharmacophore Validation. Generally speaking, a pharmacophore model should explain how structurally diverse ligands can bind to a common receptor site. Important attributes of an overlay include a high shape match because we assume that all ligands bind in the same pocket, a low 2D similarity to ensure a wide range of molecular diversity, and a high fitting of similar steric and electronic features to detect which part of each molecule is involved in establishing key interactions. While shape similarity is not the only determinant of whether two ligands will bind into the same site, it is a very parsimonious descriptor, being based only on atomic volumes. If an overlay set has high shape similarity, it is unlikely that more detailed treatment of the pharmacophore is required. Any pharmacophore elucidation method should be able to reproduce overlays with high shape similarity, if only because they will encode shape in some way. On the basis of the protocol described in the Materials and Methods section, we found that ligands from uricase are the best overlaid. They shared a high shape similarity and color score but also a fairly low fingerprint similarity. However, it should be noted that as a consequence of the nature of 2D fingerprints, such as foyfi, for such small molecules, the fingerprint similarity can be low even when the structural differences in the scaffold are minimal. Figure 6 shows the target overlay in the binding site of the reference structure. Common features are well superimposed although embedded in different 2D scaffolds. Uricase is an enzyme that catalyzes the oxidation of uric acid to allantoin, although its coding gene is inactivated in humans. Consequently, uricase has been formulated as a protein drug to overcome severe disorders induced by uric acid accumulation.

Another example of a good overlay was the case of the estrogen receptor beta (ERβ), where 18 diverse ligands were nicely superimposed (Figure 7). It is an interesting drug target because ERβ-specific agonists are used in colon cancer therapy. The data set shared four pharmacophoric points: two hydrophobic features that were full parmacophoric points and two donor/acceptor groups. In the latter case, the point labeled as A1/D1 in Figure 7B was related to a hydroxyl group present in all but one ligand, while the point labeled as A2/D2 in Figure 7B was again a hydroxyl group but was absent in four ligands.

Interestingly, the nine targets post-processed through the contact analysis all appear as good overlays apart from thrombin and glycogen phosphorylase that are both ranked in lower positions due to the high 2D similarity of the ligands in the data sets. A retrospective analysis of the original sets demonstrated that the new smaller data set has improved quality.

This analysis has rated renin, caspase 3, and phospholipase A2 as the poorest overlays for different reasons. The first was penalized by the high 2D-fingerprint similarity, otherwise it was a very good overlay (Figure 8A). The second showed low shape matching, and in fact, although occupying the same protein pocket, these ligands show very different binding modes (Figure 8B). The third contained 16 ligands with very few features in common and hence a low color score (Figure 8C).

Table 3 contains the list of molecular overlays annotated with the 2D fingerprint similarity, average shape similarity, average color score, and features clustering score, sorted by the consensus maximum described above. Interestingly, 48 targets are penalized by the fingerprint similarity, 35 targets by the average shape similarity, and 38 targets by the average color score. There is no bias toward one of these descriptors in our assessment of the overlays.

CONCLUSIONS

A prerequisite for the validation of programs such as those for the elucidation of pharmacophores is a set of overlaid protein−ligand complexes where the active sites of the proteins have been overlaid in such a manner that the ligands are in a sensible common reference frame. We have presented an algorithm for performing this overlay based on PROSITE motifs where available and, hence, derived a new validation set for pharmacophore applications. We have applied the procedure to 121 sets of complexes and hereby make them available for the benefit of others in the field. Within each set, we have put the molecules into a sensible charge/tautomer state and checked for the existence of allosteric sites. We proposed a rational way of reducing the number of ligands in highly populated sets based on common contact networks. Every set has been graded as good or poor for pharmacopore validation on the basis of different parameters such as the internal shape similarity, the 2D fingerprint similarity, or the color score similarity. As a proof of the utility of this benchmarking set, we are preparing a follow up manuscript where we present validation of a recently published overlay program.

ASSOCIATED CONTENT

Supporting Information

List of targets surviving the filtering process and notes about questionable conformations observed for a few PDB structures. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ilenia.giangreco@astrazeneca.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to members, past and present, of the CCDC, in particular, Juliette Pradon, Jason Cole, John Liebeschuetz, and Robin Taylor, for discussions of the pharmacophore validation process.

REFERENCES

(1) Yang, S. Y. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discovery Today* **2010**, *15*, 444−450.

(2) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in drug design–A review. *Curr. Top. Med. Chem.* **2010**, *10*, 95−115.

(3) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des* **2000**, *14*, 215−232.

(4) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539−558.

(5) Klabunde, T.; Giegerich, C.; Evers, A. MARS: Computing three-dimensional alignments for multiple ligands using pairwise similarities. *J. Chem. Inf. Model.* **2012**, *52*, 2022−2030.

(6) Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. GRID-based three-dimensional pharmacophores I: FLAPpharm, a novel approach

for pharmacophore elucidation. *J. Chem. Inf. Model.* **2012**, *52*, 2587−2598.
(7) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653−681.
(8) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.
(9) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(10) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.
(11) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647−671.
(12) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des.* **2006**, *67*, 370−372.
(13) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567−587.
(14) Cho, S. J.; Sun, Y. FLAME: A program to flexibly align molecules. *J. Chem. Inf. Model.* **2006**, *46*, 298−306.
(15) Jones, G. GAPE: An improved genetic algorithm for pharmacophore elucidation. *J. Chem. Inf. Model.* **2010**, *50*, 2001−2018.
(16) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric accuracy of three-dimensional molecular overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996−2002.
(17) Taylor, R.; Cole, J. C.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Korb, O. Development and validation of an improved algorithm for overlaying flexible molecules. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 451−472.
(18) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein−ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.
(19) Cross, S.; Ortuso, F.; Baroni, M.; Costa, G.; Distinto, S.; Moraca, F.; Alcaro, S.; Cruciani, G. GRID-based three-dimensional pharmacophores II: PharmBench, a benchmark data set for evaluating pharmacophore elucidation methods. *J. Chem. Inf. Model.* **2012**, *52*, 2599−2608.
(20) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(21) Leo, A. J. *CLOGP*, version 4.3; Daylight Chemical Information Systems: Irvine, CA, 1991.
(22) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(23) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115−119.
(24) Sigrist, C. J.; Cerutti, L.; Hulo, N.; Gattiker, A.; Falquet, L.; Pagni, M.; Bairoch, A.; Bucher, P. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* **2002**, *3*, 265−274.
(25) *OEChem*, version 1.7.4; OpenEye Scientific Software: Santa Fe, NM, 2012.
(26) *Maestro*, version 9.2.112; Schröedinger, LLC: New York.
(27) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des* **2007**, *21*, 681−691.
(28) *MATLAB*, version 7; The MathWorks, Inc.: Natick, MA.
(29) Bondi, A. van der Waals volumes and radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(30) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(31) Blomberg, N.; Cosgrove, D. A.; Kenny, P. W.; Kolmodin, K. Design of compound libraries for fragment screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 513−525.
(32) Muller, S.; Filippakopoulos, P.; Knapp, S. Bromodomains as therapeutic targets. *Expert Rev. Mol. Med.* **2011**, *13*, e29.
(33) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein−ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.
(34) Kinoshita, T.; Nakanishi, I.; Terasaka, T.; Kuno, M.; Seki, N.; Warizaya, M.; Matsumura, H.; Inoue, T.; Takano, K.; Adachi, H.; Mori, Y.; Fujii, T. Structural basis of compound recognition by adenosine deaminase. *Biochemistry* **2005**, *44*, 10562−10569.