JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

GRID-Based Three-Dimensional Pharmacophores II: PharmBench, a Benchmark Data Set for Evaluating Pharmacophore Elucidation Methods

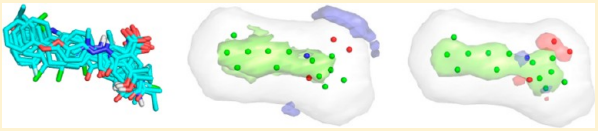
Simon Cross,*[,†] Francesco Ortuso,[‡] Massimo Baroni,[†] Giosuè Costa,[‡] Simona Distinto,[‡] Federica Moraca,[‡] Stefano Alcaro,[‡] and Gabriele Cruciani[§]

[†]Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, United Kingdom

[‡]Laboratory of Computational Medicinal Chemistry, Department of "Scienze della Salute", University "Magna Græcia" of Catanzaro, Viale Europa, Loc. Germaneto, 88100 Catanzaro, Italy

[§]Laboratory for Chemometrics and Cheminformatics, Chemistry Department, University of Perugia, Via Elce di sotto 10, I-06123 Perugia, Italy

ABSTRACT: To date, published pharmacophore elucidation approaches typically use a handful of data sets for validation: here, we have assembled a data set for 81 targets, containing 960 ligands aligned using their cocrystallized protein targets, to provide the experimental "gold standard". The two-dimensional structures are also assembled to remove conformational bias; an ideal method would be able to take these structures as input, find the common features, and reproduce the bioactive conformations and their alignments to correspond with the X-ray-determined gold standard alignments. Here we present this data set and describe three objective measures to evaluate performance: the ability to identify the bioactive conformation, the ability to identify and correctly align this conformation for 50% of the molecules in each data set, and the pharmacophoric field similarity. We have applied this validation methodology to our pharmacophore elucidation method FLAPpharm, that is published in the first paper of this series and discuss the limitations of the data set and objective success criteria. Starting from two-dimensional structures and producing unbiased models, FLAPpharm was able to identify the bioactive conformations for 67% of the ligands and also to produce successful models according to the second metric for 67% of the Pharmbench data sets. Inspection of the unsuccessful models highlighted the limitation of this root mean square (rms)-derived metric, since many were found to be pharmacophorically reasonable, increasing the overall success rate to 83%. The PharmBench data set is available at http://www.moldiscovery.com/PharmBench, along with a web service to enable users to score model alignments coming from external methods in the same way that we have presented here and, therefore, establishes a pharmacophore elucidation benchmark data set available to be used by the community.



INTRODUCTION

In part I of this paper,[1] we introduced a new pharmacophore elucidation approach FLAPpharm, that is based on GRID molecular interaction fields (MIFs),[2] and also the derivative FLAP approach[3] for molecular alignment. FLAPpharm attempts to find the best common superposition of active ligands, without being dependent on any one of them as a template, and by using each ligand's MIFs to drive the alignment. The resulting alignment model is then used to derive the common pharmacophore, unlike classical pharmacophore approaches where rule-based features are extracted, common features found, and then these are used to drive the alignment. The common pharmacophore extracted by FLAPpharm is a pharmacophoric "pseudomolecule", which consists of common pharmacophoric interaction fields (PIFs), common atom-centered pharmacophoric pseudofields (pseudoPIFs), and common pharmacophoric points at the centroid of these pseudoPIFs. These three entities directly correspond to the MIFs, pseudoMIFs, and atoms in a ligand; therefore, the FLAPpharm pseudomolecule can be used just as another molecule in FLAP; it can be used as a template for ligand-based virtual screening or it can be docked into a receptor by FLAP to help validate or disprove the hypothesis. In part I,[1] we validated this approach in several ways. First (and primarily), we used the data sets previously reported by Patel et al.[4] to gauge how successful the method was at reproducing the target pharmacophore, and in all cases obtained extremely satisfying results. Second, we used the DUD data set[5] to test how discriminatory the method was in terms of virtual screening, after automatically constructing pharmacophore hypotheses from the DUD chemotype cluster centroids identified by Good;[6] on average, the FLAPpharm approach (starting without three-dimensional information about the ligands or target) performed better than any of the other approaches we previously tested[7] (including those using known three-dimensional information about the ligands and target). Finally, we illustrated the use of a FLAPpharm model in explaining the alternative binding modes available to factor Xa, by docking the pharmacophoric pseudomolecule into the receptor. The

Received: March 22, 2012
Published: September 12, 2012

   dx.doi.org/10.1021/ci300154n | *J. Chem. Inf. Model.* 2012, 52, 2599−2608

"chloro-binding" mode, which was a significant discovery in that it allows the design of neutral and hence potentially more pharmacokinetically desirable compounds,[8] was predicted as the third docking pose. Therefore we believe that FLAPpharm shows great promise and, as discussed in part I, does not suffer from the disadvantages of the classical feature-based methods.

Among the most widely known of these feature-based methods are Catalyst/HipHop,[9] GASP,[10] and the more recently described LigandScout,[11] PharmID,[12] and PHASE.[13] The first of these (Catalyst/HipHop) was first published in 1996 and validated using three data sets. The second was first published in 1995 (GASP) and validated using eight data sets. PharmID (2006) was validated using two data sets. LigandScout (2005) differs in that it was introduced as a structure-based pharmacophore method and two validation examples were used, although a follow-up paper examined five case studies.[14] The current version of LigandScout also includes a ligand-based approach.[15]

Recently another computational approach based on the generation of pharmacophore models by GRID maps of complexes (GBPM) was described by some of us, validated in two systems,[16] and subsequently successfully applied to HIV-1 reverse trascriptase case studies.[17] PHASE (2006) was validated using the Patel data set[4] that we also examined in part I of this series. Typically then, only a handful of data sets are used to validate the algorithms, and even these data sets are not consistent across the different publications. Leach et al. also point out that pharmacophore elucidation programs are occasionally validated on series of ligands that any competent modeler could mentally overlay in a few seconds.[18] A step forward in this area is described by Jones when evaluating GAPE,[19] a follow-up to the GASP method, which uses 13 target sets for validation, and also describes an objective methodology for evaluating the quality of the alignments.

In the area of virtual screening, in recent years a number of publications have appeared to address the similar problem of validating these methods. The DUD data set is an admirable attempt to provide a curated benchmark data set with known active ligands against a large number of targets, with more relevant decoys, so as to avoid bias (for example where decoys could be easily distinguished from actives by trivial metrics such as molecular weight). In our opinion, there is clearly a need for a similar benchmark data set in the area of pharmacophore elucidation, and our aim with this publication is to provide such a data set, some objective metrics to evaluate performance, and describe how our own method performs.

■ ASSEMBLING THE PHARMBENCH DATA SET

To validate pharmacophore elucidation methods, for each data set the correct answer must be available to compare with the predicted hypotheses. The first limitation is therefore that experimental structural data must be available, and we added the restriction that the ligands themselves must be pharmaceutically relevant. As a first step, we used the online DrugPort[20] resource as a starting point, which lists the protein targets for approved drugs and nutraceuticals. Therefore any structural target in the data set would be one for which there is an approved drug or nutraceutical. The full list of ~1600 targets was downloaded, and the list was filtered to remove targets with fewer than four structures and having at least one drug on the market, to leave 334 targets. The UniProtKB accession number for each target was then used to query the protein data bank,[21] with the additional requirement that the structure must contain a ligand. In this way the pdb structure files were obtained for each target. The Ligand Expo database[22] of all small molecule structures present in the protein data bank was downloaded, and for each of the targets in our data set, the small molecules present in the files were analyzed and the target structure was filtered if the ligand did not match the following constraints: molecular weight greater than 100 and less than 500; number of rotatable bonds less than 12; number of hydrogen bond donors less than or equal to 5; number of hydrogen bond acceptors less than or equal to 10; logP less than 5.0. The remaining target structures therefore contained small molecule ligands that were "drug-like". Small molecules were also filtered if their incidence in the Ligand Expo database was higher than 20, as an attempt to automatically remove cofactors and other small molecules that are not relevant. For each target set, the structure files were also removed if the resolution of the PDB structure was greater than 2.5 Å, to leave only higher resolution structures. As an additional filter, we removed structures that did not have electron density deposited at the Uppsala EDS server[23] to ensure that if needed, the original experimental data could be queried. As a final filter, each target was filtered unless the number of unique ligands was greater than or equal to three. This left a total of 94 targets to analyze in more detail by manual inspection. To do this, for each target, the protein structures were first aligned using the CE algorithm as implemented in PyMOL,[24] the small molecules extracted, and the atom-typing from the Ligand Expo database mapped onto the aligned PDB structures of the ligands to provide the aligned ligand sets for each target. The ligands were then corrected by hand, with particular respect to tautomerism and protonation. Duplicates were removed, and any other structures where there were problems with chemistry or the alignment. Given that some ligands had been removed, we filtered targets that no longer had at least three unique ligands. Finally, we added the five targets from the Patel data set, and a set of high resolution DNA major groove binders given our interest in this area. This yielded a final data set of 81 targets containing 960 ligands, taken from high resolution crystallography data, and aligned by the target protein structure, which we will henceforth refer to as Pharmbench. All ligands were also converted to their equivalent two-dimensional structures using the dbtranslate utility in SYBYL X1.3,[25] to remove any bias when using them as input to a pharmacophore elucidation program. Our aim is that this will become a publicly available resource, downloadable at www.moldiscovery.com/PharmBench, to be maintained and updated over time as new structural data becomes available, to enable molecular alignment and pharmacophore elucidation approaches to be compared. Naturally, being version 1.0 of the benchmark set, there are some limitations. The most obvious of these is that currently no diversity filtering has been applied to the PharmBench data sets; therefore, it is possible for there to be "easy" cases where simple analogues are present. However, at this stage, we decided to retain all of the experimental information that passed our preliminary filters and try to set up an objective measure of success that can be used by the community.

■ GENERATING THE HYPOTHESES

We have already described the FLAPpharm method in part I of this paper;[1] however, some additional options were tested in this work. The first of these concerns the molecular alignment when constructing the model; the alignment is performed using quadruplets formed from combinations of the atoms in the

Table 1. Pharmacophore Elucidation Performance by FLAPpharm on the PharmBench Data Set[a]

a

Target Name	UniProtKB ID*	nMols	Bioactive conf evaluation: 2D + 30 confs % mols < 2.0 Å	AlignScore Alignment evaluation: Xray confs	AlignScore Alignment evaluation: 2D + 30 confs	Pharmacophoric similarity evaluation: xray confs PIF similarity to Gold Standard	Pharmacophoric similarity evaluation: 2D + 30 confs PIF similarity to Gold Standard
Oxidoreductases							
aldose reductase	P15121	24	75	88	38	1.00	0.75
monoamine oxidase	P27338	18	94	50	39	0.96	1.00
amine oxidase	P46881	16	44	75	44	0.92	0.75
dihydrofolate reductase	P00374	16	88	88	44	1.00	1.00
dihydroorotate dehydrogenase	Q02127	13	100	92	100	1.00	1.00
enoyl-reductase	P0A5Y6	10	100	100	70	1.00	0.54
horse alcohol dehydrogenase	P00327	9	89	100	89	0.51	0.33
cytochrome P450 2A6	P11509	7	100	100	86	0.97	0.59
dihydrofolate reductase	Patel_DHFR	6	83	100	50	1.00	1.00
human alcohol dehydrogenase	P00325	4	100	75	75	0.67	0.58
cytochrome P450cam	P00183	4	75	75	75	0.63	0.69
cytochrome P450 51	P0A512	4	75	50	25	0.87	0.67
human phenylalanine hydroxylase	P00439	4	100	100	100	0.94	0.87
cytochrome P450 2B4	P00178	3	100	67	67	0.65	0.50
dihydroorotate dehydrogenase	Q08210	3	100	100	100	1.00	1.00
Transferases - General							
HIV reverse transcriptase	Patel_HIVRT	10	80	60	30	1.00	1.00
purine nucleoside phosphorylase	P0ABP8	10	100	100	100	1.00	0.63
farnesyl pyrophosphate synthetase	P14324	7	100	100	100	1.00	0.84
uridine phosphorylase	P12758	6	100	100	50	1.00	0.51
branched chain amino acid aminotransferase	P54687	3	100	100	100	0.53	0.49
glutathione S-transferase	P09211	3	100	67	67	0.54	0.50
thymidine phosphorylase	P19971	3	100	100	100	1.00	1.00
Transferases - Kinases							
glycogen synthase kinase-3 beta	P49841	15	80	60	27	1.00	1.00
thymidine kinase	P03176	12	100	100	75	1.00	0.68
tyrosine protein kinase LCK	P06239	9	100	89	22	1.00	1.00
vascular endothelial growth factor receptor 2	P35968	8	62	100	50	1.00	0.61
mitogen-activated protein kinase	P28482	6	83	100	33	0.86	0.62
cyclin-dependent kinase 2	Patel_CDK2	6	100	67	50	1.00	0.95
tyrosine kinase ABL1	P00519	6	83	83	83	1.00	1.00
CFMS tyosine kinase	P07333	4	100	75	100	1.00	1.00
protein kinase B	P31749	4	100	100	100	1.00	0.86
fibroblast growth factor receptor 1	P11362	3	100	67	67	0.94	0.75
Hydrolases - General							
dipeptidyl peptidase IV	P27487	30	87	80	43	0.67	0.67
acetylcholinesterase	P04058	19	79	21	16	0.82	0.68
phosphodiesterase 4B	Q07343	15	80	40	40	0.69	0.67
protein-tyrosine-phosphatase 1B	P18031	13	62	85	62	0.91	0.75
phosphodiesterase 4D	Q08499	11	64	55	36	1.00	0.97
ampc beta-lactamase	P00811	10	80	50	40	1.00	1.00
ribonuclase A	P61823	7	100	100	100	1.00	0.88
renin	P00797	7	86	86	86	1.00	0.92
glutamate carboxypeptidase II	Q04609	6	83	100	67	1.00	0.67
caspase-3	P42574	6	83	83	67	1.00	0.49
epoxide hydrolase 2	P34913	6	100	83	83	0.98	0.87
phosphodiesterase	O76074	6	67	67	17	0.91	0.85
phospholipase A2	P59071	3	100	67	67	0.71	0.80
histone deacetylase-8	Q9BY41	3	100	100	100	1.00	0.81

b

Target Name	UniProtKB ID*	nMols	Bioactive conf evaluation: 2D + 30 confs % mols < 2.0 Å	AlignScore Alignment evaluation: Xray confs % mols < 2.0 Å	AlignScore Alignment evaluation: 2D + 30 confs % mols < 2.0 Å	Pharmacophoric similarity evaluation: xray confs PIF similarity to Gold Standard	Pharmacophoric similarity evaluation: 2D + 30 confs PIF similarity to Gold Standard
Hydrolases - Serine Proteases							
beta-trypsin	P00760	83	100	43	16	0.36	0.37
human alpha thrombin	P00734	41	68	54	24	0.62	0.50
human factor Xa	P00742	26	69	65	19	0.51	0.50
urokinase-type plasminogen activator	P00749	14	100	79	50	0.69	0.45
thrombin	Patel_THROMB	7	57	100	57	0.95	0.50
factor VII	P13726	6	50	100	50	0.98	0.69
factor VIIA	P08709	5	100	100	40	1.00	0.63
factor XI	P03951	4	100	100	75	1.00	0.56
Hydrolases - Metalloproteases							
matrix metalloprotease-12	P39900	15	93	87	60	1.00	0.62
thermolysin	Patel_THERM	6	67	100	33	1.00	0.56
matrix metalloprotease-9	P14780	3	100	100	67	0.94	0.46
Lyases							
human carbonic anhydrase II	P00918	84	88	67	33	0.75	0.54
human carbonic anhydrase I	P00915	7	100	100	100	1.00	0.93
human carbonic anhydrase IV	P22748	3	100	100	67	1.00	0.56
Nuclear Receptors							
estrogen receptor	P03372	38	84	71	53	1.00	1.00
PPAR-gamma	P37231	22	64	41	9	0.69	0.55
androgen receptor	P10275	16	69	50	56	0.74	0.46
vitamin D nuclear receptor	P11473	13	92	100	85	1.00	0.63
PPAR-alpha	Q07869	6	83	83	17	1.00	0.68
retinoic acid receptor RXR-alpha	P19793	6	100	83	83	1.00	0.80
mineralocorticoid receptor	P08235	6	100	100	100	1.00	1.00
estrogen related receptor gamma	P62508	5	100	80	100	1.00	0.94
estrogen-receptor beta	Q92731	5	100	80	80	1.00	0.96
glucocorticoid receptor	P04150	4	100	75	50	0.65	0.46
retinoic acid receptor gamma 2	P13631	4	100	100	100	1.00	0.73
PPAR-delta	Q03181	3	67	100	33	1.00	0.67
thyroid hormone receptor beta-1	P10828	3	100	100	100	1.00	0.80
Other							
HSP 90 alpha	P07900	57	75	56	21	0.75	0.75
glutamate receptor 2	P19491	14	57	36	14	0.50	0.50
glutamate receptor 5	P39086	6	100	100	100	1.00	1.00
triosephosphate isomerase	P04789	3	100	100	100	1.00	0.81
DNA	DNAFED	12	75	67	67	1.00	0.83
transthyretin	P02766	28	96	46	46	1.00	0.97
androgen-binding protein	P04278	4	100	75	75	0.85	0.78
retinoic acid binding protein type II	P29373	3	100	100	100	1.00	0.83

[a]The table shows the target name, UniProtKb IDs for each target (with the addition of pseudocodes for the Patel data sets and our DNA minor groove binder data set), the number of molecules aligned for that target, and the results of the three benchmark metrics when building models from the X-ray conformations or from the 2D input structures and up to 30 generated conformations. The second AlignScore evaluation metric is colour coded to indicate successful models (50% or more molecules aligned with an rmsd of <2.0 Å, green) and unsuccessful models (<50% molecules aligned with an rmsd of <2.0 Å, pink). See text for details.

dx.doi.org/10.1021/ci300154n | J. Chem. Inf. Model. 2012, 52, 2599−2608

molecules, and the subsequent molecular alignments are then scored using the MIF similarity. Naturally there are many quadruplets present in each conformer of each ligand and template molecule; the *best* option allows the comparison of more geometrically dissimilar quadruplets, the *fast* option filters all but only the most geometrically similar quadruplets, and the default option (termed *normal* in the results below) is a compromise between them. The second option is called *quickmodel*, and this bypasses the full tree search, attempting to find the correct branch of the tree to search by ranking each molecule in the data set by its ability to be a good template for the other molecules, according to a proprietary algorithm. The third option tested is called *msi* and involves a postprocessing step to refine the model. Molecules are realigned to their original template; however, an additional local conformational search is performed, to ensure that common substructural features are well-aligned. Hypotheses were also constructed starting from the X-ray conformations (rigid alignment) and also starting from the two-dimensional input structures and generating up to 30 lowest energy conformations (unbiased model). For each of these 24 combinations, the top-ranked 5 models were retained for analysis, yielding 120 models for each of the 81 data sets (9720 models in total).

VALIDATING PHARMACOPHORE ELUCIDATION APPROACHES

One of the key problems when validating pharmacophore elucidation methods is how to define success.[19,26] Clearly, given the experimentally derived alignments in PharmBench, we have available a gold standard; any approach should at best be able to reproduce identical alignments to those given by the experimental data, starting from the two-dimensional structures and without any prior knowledge of the three-dimensional ideal solution. The common features present in the gold standard should also be identified by the algorithm. These aspects are also discussed by Jones[19] during the validation of GAPE, although a subjective measure was also applied to account for potential misleading flaws in the objective measures. For example, the set of overlays predicted by the algorithm may be in a completely different frame of reference to the experimentally derived gold standard, and these overlays need to be aligned according to some consistent method to compare root mean squared deviation (rmsd) values across the model compounds. Each molecule in turn can be aligned to its experimental counterpart to determine if the bioactive conformation (or close to it) has been found; however, this gives no indication about the hypothesis overlay quality. A molecule with a poor overall rmsd to its experimental conformer may fit well to the hypothesis but contain a flexible solvent accessible "tail" that is not constrained by the model and therefore "incorrectly" predicted. Identifying common features is also problematic; different programs may recognize features differently, and in our case there are no individual features but only pharmacophoric interaction fields. We therefore decided to introduce three objective measures:

1. **Bioactive conformer evaluation**. To measure how well the software was able to predict the bioactive conformation, each predicted molecule conformation was aligned to its experimental counterpart (with equivalent atoms assigned using a subgraph matching algorithm, followed by least-squares fitting) and the heavy atom rmsd calculated. Success was measured if the rmsd was <2.0 Å.
2. **AlignScore alignment evaluation**. To measure the alignment quality, we have adopted the method proposed by Jones;[19] the model is aligned to the experimental gold standard alignments using an iterative approach, whereby various combinations of molecules are used to align the model to their molecules in the X-ray alignments. Exhaustively considering all combinations of molecules to perform the alignment would be prohibitively time-consuming for large data sets; hence the following iterative procedure was used. The first molecule is added to the *FittingStructures* set, and the atoms in this set are used to perform a least-squares fit to their equivalents in the X-ray alignments (all other molecules are rototranslated in the same way). The rmsd is evaluated for all molecules in this alignment compared to the X-ray alignment. If the rmsd for the *FittingStructures* atoms is >2.0 Å, this molecule is removed from further consideration and classed as incorrectly predicted. If the rmsd is <2.0 Å, the molecule is considered as correctly predicted and added to the FittingStructures set. From the remaining molecules in this model alignment, the molecule with the lowest rmsd is then added to the FittingStructures set, and the process is repeated. The final result is the iteration with the largest number of structures in the FittingStructures set, and in the result of a tie, the set with the lowest rmsd is chosen. Henceforth, we will refer to this metric as "AlignScore". The objective measure of success is whether at least 50% of the molecules are contained within the best set (i.e., AlignScore $\geq$ 50).
3. **Pharmacophoric similarity evaluation**. To measure how similar the hypothesis was to the gold standard in terms of pharmacophoric similarity, the pharmacophore (a pharmacophoric pseudomolecule) was perceived for the experimental alignments. The pharmacophore models were then aligned to this gold standard, using FLAP and their PIFs to optimize the alignment. The resulting field similarities provide a quantitative measure of how similar the models are to the experimental alignments.

FLAPPHARM PERFORMANCE ON THE PHARMBENCH DATA SET

For each of the 81 PharmBench data sets, FLAPpharm was used with the options described above to produce 60 "rigid alignments" and 60 "unbiased models". The best of the rigid alignments and unbiased models according to the PIF similarity measure was retained for subsequent analysis. Table 1a and b report the evaluation data for the 81 data sets, grouped by target class, and for both the rigid alignments and unbiased models. In this section, we will analyze the overall performance according to the different evaluation metrics, examine specific cases, and discuss the merits of the objective evaluation metrics.

IDENTIFICATION OF THE BIOACTIVE CONFORMATION

The first of the evaluation metrics to be analyzed is the rmsd of the molecules in the predicted models, compared to their experimental X-ray conformers, and tests solely whether the method has been able to determine the bioactive conformation, or close to it. Figure 1 shows the average percentage of

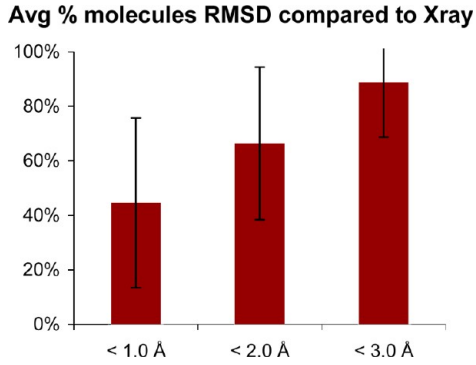


Figure 1. Finding the bioactive conformations. from the best models predicted by FLAPpharm (starting from two-dimensional (2D) input structures), the average percentage of molecules lower than a threshold is reported. Error bars illustrate the population standard deviation.

molecules lower than an rmsd threshold of 1, 2, and 3 Å, for the best models starting from the two-dimensional input files. It can be seen that 65% of the 960 input molecules are predicted satisfactorily (rmsd < 2 Å). This compares with the 91% success found by validating the conformer generator; in other words, of the 91% of molecules that the conformer generator is able to produce close to the bioactive conformation, the models are selecting these a priori with 71% success. Examining the individual data sets, 67 of the 81 (83%) have models with greater than 75% of the ligands' bioactive conformations predicted and identified successfully, and only for one case are fewer than 50% identified (P46881). In general, we interpret this as fairly good performance by the method in identifying the bioactive conformations.

IDENTIFICATION OF THE CORRECT ALIGNMENT

The second AlignScore metric evaluated the alignment quality of the model to the experimentally derived X-ray "gold standard" structure alignment, using an iterative approach to search for a superposition which yielded the largest number of molecules where their predicted conformation and alignment compared to the reference was <2.0 Å.

A successful model is one where 50% or more of the molecules are identified by this alignment as being <2.0 Å compared to the gold standard (AlignScore ≥ 50), as reported by Jones.[19] It should be noted that it is an iterative (and not exhaustive) approach, it is therefore possible that more optimal superpositions could be found, i.e. that the real quality of the models is potentially higher than that identified and reported below.

For the rigid alignment, a successful model was produced for 93% of the data sets, reinforcing our conclusion from part I that MIFs demonstrate excellent performance when used to align molecules. For the unbiased models, this success rate drops to 67%, which is unsurprising given that the method has to cope with selecting the correct conformation from many, in addition to identifying the correct alignment. It also appears from Table 1a and b that the failures according to this metric appear higher for those data sets with larger numbers of compounds. There are several reasons why this might be the case: First, these data sets are inherently more complex so finding the bioactive conformation and aligning it correctly is more difficult. Second, the AlignScore metric is itself iterative and not exhaustive; therefore with more complexity, the AlignScore is likely to be underestimated. With larger numbers of compounds, there is also a higher chance that there will be structural analogues present, and this may be more detrimental to the model since these analogues may have many conformers that score and align well with each other, yet not be the bioactive conformation.

Below we will examine in more detail those data sets for which a good model was not identified according to this metric.

EVALUATION OF THE PHARMACOPHORE INTERACTION FIELD SIMILARITY

The third metric evaluated was the PIF similarity between each of the 9720 models and the gold standard pharmacophoric pseudomolecule percieved from the X-ray structure alignments. The similarity is measured in terms of four PIFs, coming from the GRID H, O, N1, and DRY probes, describing shape, hydrogen-bond acceptor, hydrogen bond donor, and hydrophobic similarity, respectively. Positive and negative charge interactions are also accounted for within the O and N1 probes. The overall similarity is simply a sum across these four values; hence, a score of 4.0 is the maximum that is achievable. If the gold standard pseudomolecule does not contain one of the fields, the similarity is reported as 0.0 for this field when comparing with a model. Therefore, we aligned the gold standard pseudomolecule to itself and normalized the model similarities by this predetermined maximum. The maximum normalized similarity is therefore 1.0.

One of our aims was to investigate the effect of the different parameters on the quality of the models; therefore, we examined the average PIF similarities for the models produced using each parameter, and these are illustrated in Figure 2. The

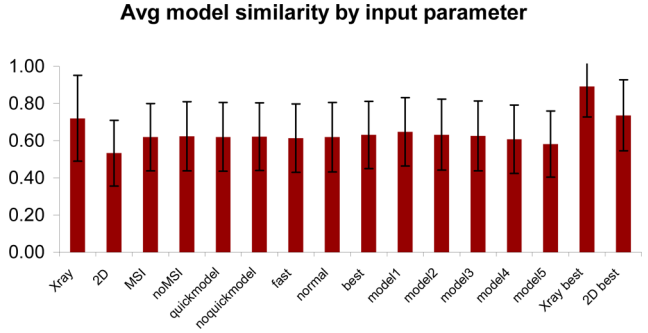


Figure 2. Normalized pharmacophoric interaction field similarities between the FLAPpharm derived hypotheses and the experimentally derived gold standard alignments, averaged across the various input parameters. Error bars illustrate the population standard deviation. See text for details.

first two columns show the average similarities produced by the rigid alignments, and the unbiased models. For the rigid alignments, the average similarity is 0.72, whereas unsurprisingly starting from the 2D input conformations the similarity is lower at 0.53. The last two columns illustrate the same result; however in this case, the best scoring model for each data set has been selected to calculate the average. Here the equivalent similarities are 0.89 and 0.74; in other words using the X-ray conformations as input, it appears the method is finding similar pharmacophoric interaction fields to those in the gold standard in the majority of cases, and the best models coming from the 2D input are not too far behind this. These values are not dissimilar to the alignment quality success found using the

dx.doi.org/10.1021/ci300154n | J. Chem. Inf. Model. 2012, 52, 2599−2608

second evaluation metric described above (93% and 67%, respectively). Given that the pharmacophoric interaction may be confined to a subset of the molecules, it is reasonable to assert that the "whole molecule rmsd measure" used by the second metric is much less forgiving than the field similarity reported here, i.e., part of each molecule may be aligned incorrectly but be irrelevant to the pharmacophore.

In terms of the various input parameters, examination of the msi and nomsi similarities appears to indicate that this option has little affect on the models; therefore, any local substructural conformational tweaking is not significantly affecting the models. The fast, normal, and best parameters also appear to have a very minor affect, implying that more geometrically dissimilar quadruplet searching is not required. As expected, the ranked models 1−5 show a natural trend from higher average scoring to lower. Most surprisingly, the proprietary quickmodel approach appears to give almost equivalent performance compared to the default more exhaustive tree search.

DISCUSSION OF SELECTED EXAMPLES

From Table 1a and b, and using the AlignScore metric of the number of molecules having an acceptable rms deviation from the X-ray structures (AlignScore ≥ 50), 93% of the rigid alignments and 67% of the unbiased models were judged to be good. In this section we will examine a couple of models that were judged to be good and, then on six of the twenty-seven cases that were judged not to be good for the unbiased models, try to understand both the limitations of FLAPpharm and also the validation metrics. For each case, we report the aligned molecules and the pharmacophoric pseudomolecule with PIFs and pseudoPIFs, for both the gold standard X-ray alignments and the best unbiased model produced by the algorithm.

Figure 3 shows the results for a "perfect" model (i.e., AlignScore = 100; PIF similarity = 1.0). The data set consists of 13 dihydroorotate dehydrogenase ligands (UniProtKB ID: Q02127). It is apparent that the alignments are much "crisper" around the carboxyl and carbonyl groups in the model; therefore, the donor field is more intense in these regions and an additional region is identified at the bottom of the image. Since the field similarity is normalized by the gold standard field, the additional region does not penalize the similarity.

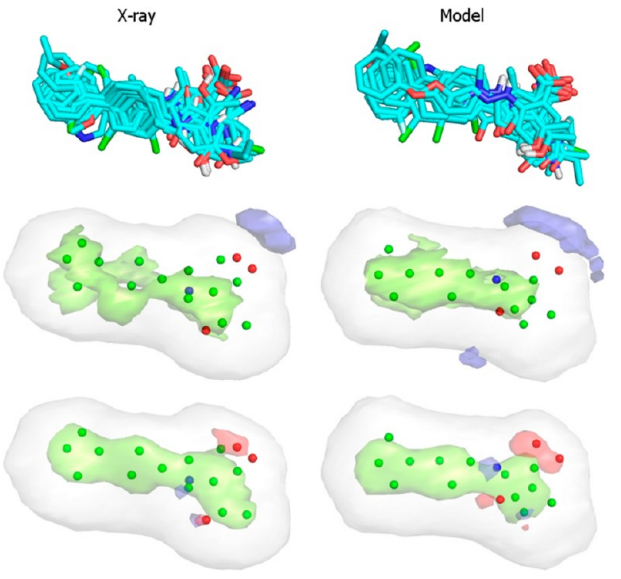


Figure 3. Q02127 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

Figure 4 shows the results for our collected set of 12 DNA minor groove binding ligands with our UniProtKB pseudocode DNAFED (AlignScore = 67, PIF similarity 0.83). In reproducing the bioactive conformations, two of the ligands have an rmsd >2.0 Å (pdb codes: 328D, 3OIE) and are therefore not found by the conformer generator. For an additional two ligands (pdb codes: 2NLM, 8BNA), the bioactive conformation is not aligned in the correct way. However, 2NLM has a high degree of field symmetry (although it is not chemically symmetrical), and in the model, it is aligned in a pharmacophorically identical manner, which highlights a limitation of using the AlignScore metric; the alignment may not be correct, but the model is still good.

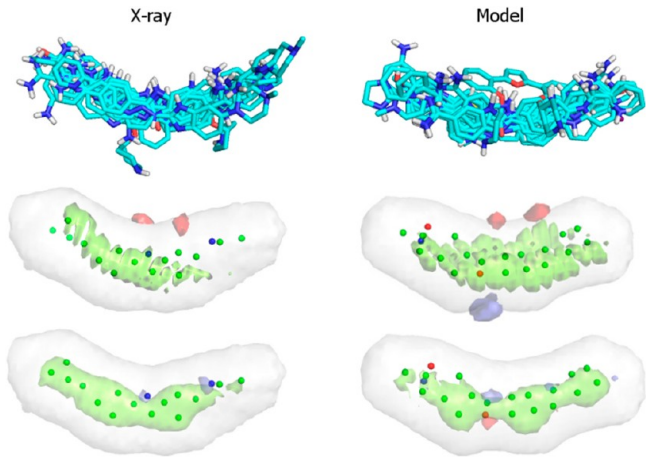


Figure 4. DNAFED ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top), and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

Figure 5 shows the results for a set of six phosphodiesterase ligands with the UniProtKB ID O76074 (AlignScore =17, PIF similarity 0.85). In this case the model similarity score is high, whereas the alignment similarity is very low. 67% of the ligands' bioactive conformations are reproduced by the conformer generator and identified by the model; however, in their alignments only one of the six is able to be aligned correctly, which is the worst possible case if the correct conformation has been found. Inspection of the model shows that it is not dissimilar; the target hydrophobic region is found in the same location and also the target donor/acceptor pair. The shape differs toward the right-hand side of the figure. The example again demonstrates that a poor alignment does not necessarily mean that the model is poor.

Figure 6 shows the results for a set of nine tyrosine kinase ligands with the UniProtKB ID P06239 (AlignScore = 22, PIF similarity 1.00). Only two of the ligands' bioactive conformations and alignments are identified correctly by the model; however, the model is perfect according to the PIF similarity. Qualitative inspection of the model compared to the

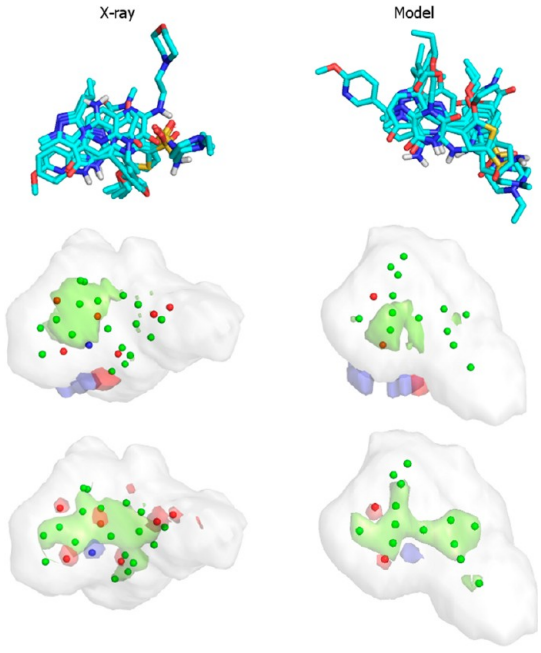


Figure 5. O76074 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

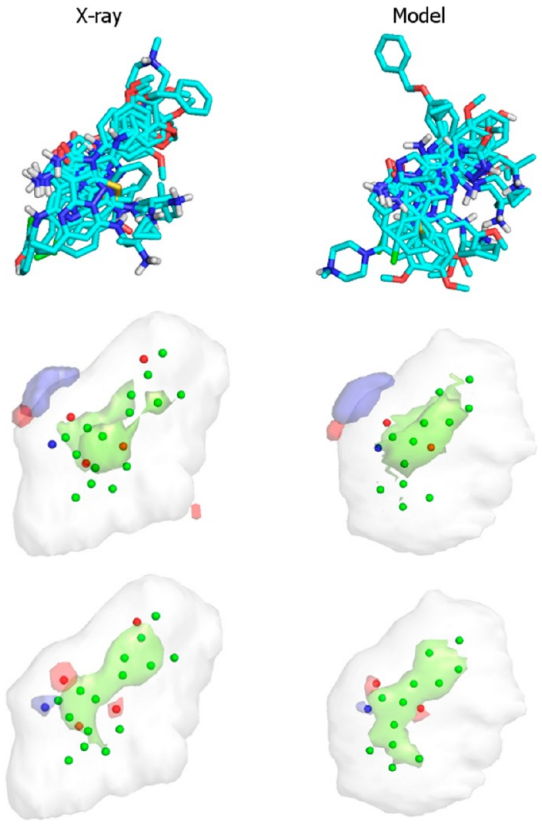


Figure 6. P06239 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

X-ray alignments shows that the typical donor/acceptor pair interacting with the kinase hinge region is identified, as well as the common hydrophobic region. This is another example where identifying the correct conformation and alignment is not necessary for correct perception of the pharmacophoric interactions.

Figure 7 shows the results for a set of four cytochrome P450 51 ligands with the UniProtKB ID P0A512 (AlignScore = 25, PIF similarity 0.67). With only 4 ligands, the AlignScore of 25 means that only one of the ligands has a conformation that has been found and aligned correctly. One of the ligands (pdb code: 2W0B) contains a *cis*-amide that is not found by the FLAPpharm conformer generator due to its high energy relative to the trans conformation. This biases the conformations in the model such that in the bottom half of the figure the model matches the X-ray alignments, but in the top half the alignments are different. The model is one of the worst cases and yet may still provide useful information.

Figure 8 shows the results for a set of 26 factor Xa ligands with the UniProtKB ID P00742 (AlignScore = 19, PIF similarity 0.50). We have already shown in part I of this series[1] a case study where a successful pharmacophoric model of factor Xa inhibitors has been built and docked into the receptor, finding the two primary binding modes in the top docking poses. For this data set, the model is incorrect, yet not unreasonable. There is primarily shape and acceptor PIF similarity between the model and the gold standard (which is likely to account for the PIF similarity of 0.5 since two of the four fields match reasonably well). It appears that the model is driven by the overlay of the sulfonamide and amide groups, whereas in the X-ray alignments the terminal rings are aligned much more cleanly. Since the rings are not aligned as cleanly in the model, the hydrophobic PIF is not identified. One aspect that makes this data set more interesting is the mixing of the "chloro-binding" mode and the primary binding mode. On the left-hand side of the X-ray alignments in the figure, charged groups such as benzamidines are overlaid with chlorophenyl moieties; the model has at least identified that these bind in the same location, even if their superposition is not optimal.

Figure 9 shows the results for a set of 57 HSP90 ligands with the UniProtKB ID P07900 (AlignScore = 21, PIF similarity 0.75). For 75% of the ligands, the bioactive conformation is reproduced; however, only for 21% of the ligands is the overlap correct. The overall PIF similarity is fairly good, and inspection of the figure shows that qualitatively the hydrophobic and acceptor PIFs have been identified, the shape is similar but not quite correct, and the donor PIF is not found.

Figure 10 shows the results for a set of three branched chain amino acid aminotransferase ligands with the UniProtKB ID P54687 (AlignScore = 100, PIF similarity 0.49). This example illustrates the opposite extreme discussed above; the alignments are perfect, but the PIF similarity is poor. First, since there are only three ligands, small differences have a greater effect on the PIFs. Second, two of the ligands are very small which means that the rmsd comparison is more forgiving. Qualitative inspection of the model shows that the interacting groups are twisted very slightly and are aligned favoring the SO_2 group slightly more than the NH_2; hence the acceptor field disappears and the donor field is more intense. There is no DRY PIF in the X-ray alignment; hence, the other three fields are contributing a greater proportion to the overall score. Therefore, completely losing the acceptor field will lose

dx.doi.org/10.1021/ci300154n | J. Chem. Inf. Model. 2012, 52, 2599−2608

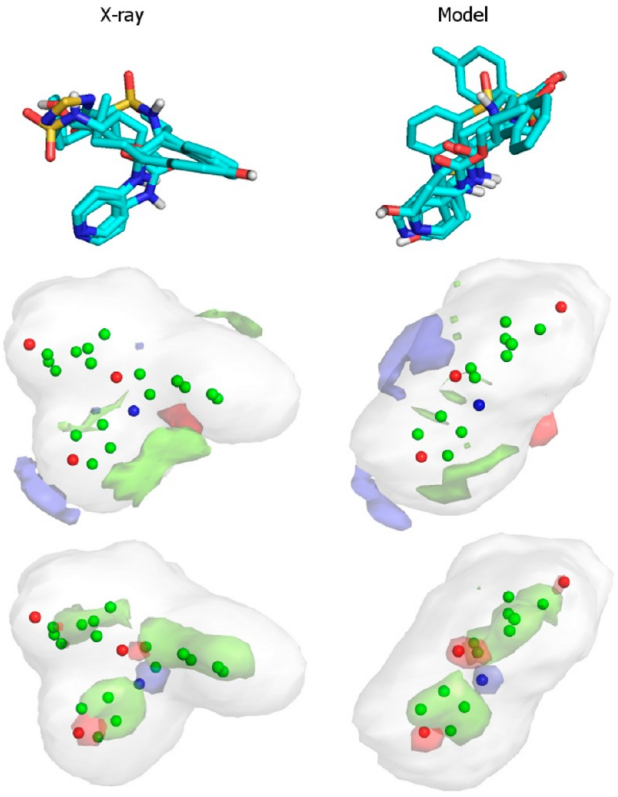


Figure 7. P0A512 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

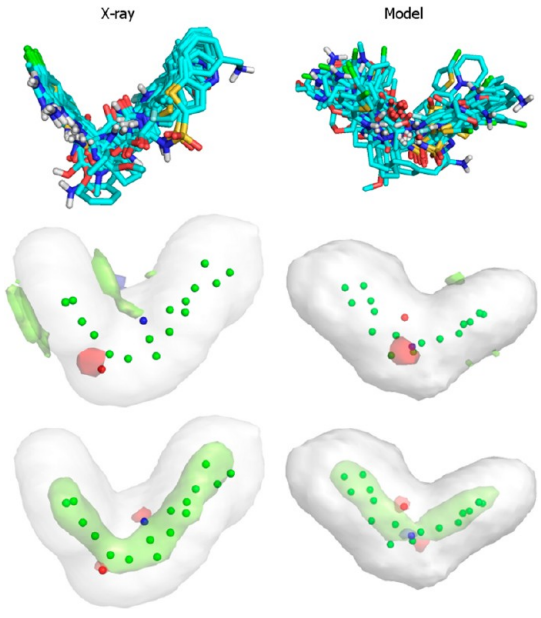


Figure 8. P00742 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

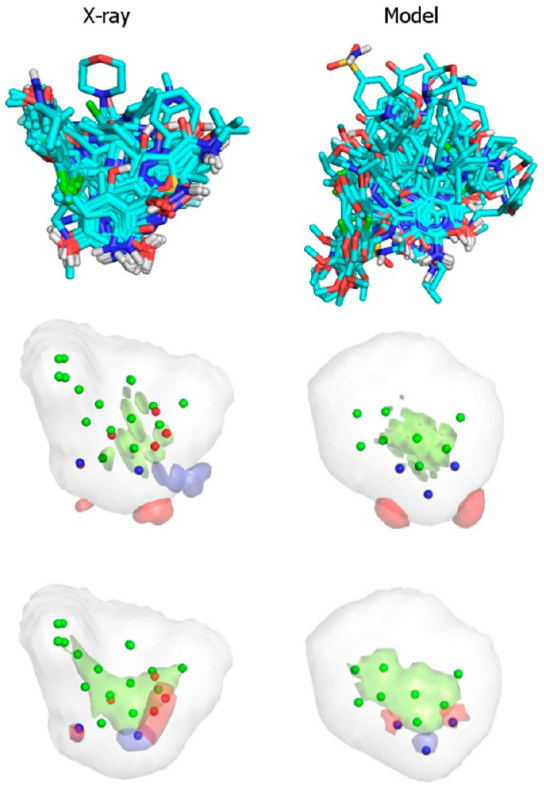


Figure 9. P07900 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

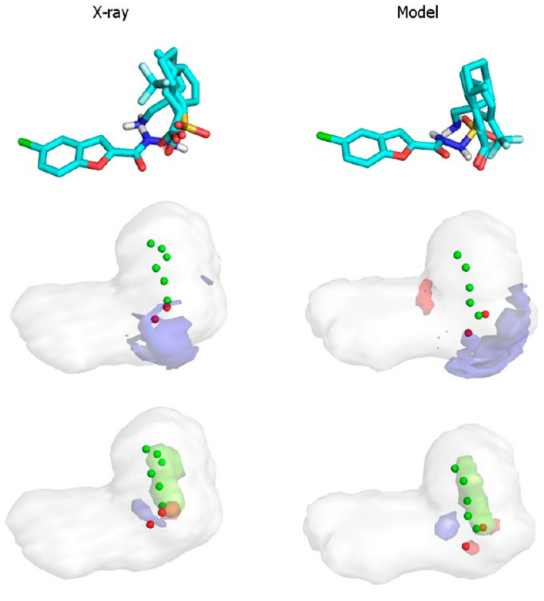


Figure 10. P54687 ligands with alignments from the X-ray structures and from the best unbiased model. The figure shows the ligand alignments (top) and the pharmacophoric pseudomolecule with PIFs (middle) and pseudoPIFs (bottom) displayed (green = hydrophobic, blue = hydrogen bond donor, red = hydrogen bond acceptor, white = shape).

approximately one-third of the similarity score, instead of one-quarter. The remainder is lost from the reoriented donor field

dx.doi.org/10.1021/ci300154n | J. Chem. Inf. Model. 2012, 52, 2599−2608

found in the model. But by inspection the pharmacophoric pseudomolecule does not look as different as the score suggests.

DISCUSSION OF THE VALIDATION METHODOLOGY

The primary aim of this paper is to report a benchmarking data set for pharmacophore elucidation, along with several quantitative measures of success; inspection of the models built using FLAPpharm has highlighted some of the potential limitations of this validation methodology. First, an individual data set may not contain enough information to enable any algorithm to find the correct model; unless some molecules exist to restrict the conformational space of the others in the set, it may be impossible to identify the bioactive conformation and alignment. For example, consider aligning a molecule to itself with *N* conformations; *N* alignment pairs will be produced which all score equivalently; therefore, the "top" few solutions may contain the bioactive conformation only by chance. Another problem is that the binding modes of the ligands may be too different, as is the case for P04058, P37231, and P19491, where one or two features may interact in the same way, but the rest of the molecule makes completely different interactions with the receptor. A final limitation of the data set is that it comes from experimental data which is also subject to error, and the gold standard alignments have been produced by aligning their crystallographic receptor atoms, which is likely to be less than optimal. This was highlighted in part I of this series,[1] where the FLAPpharm alignment of the X-ray conformations of the thermolysin ligands from the Patel data set appeared to be much crisper than the gold standard alignments.

Second, it is intuitive that identification of the bioactive conformation and correct alignment will result in the correct identification of the common pharmacophore; however, failure to find the correct conformation and alignment does not necessarily mean that the identified pharmacophore is incorrect, as the examples above illustrate. This may be due to symmetry in a ligand's interaction fields, or the fact that the pharmacophoric interaction is limited to a specific region in the ligand, effectively allowing the rest of the ligand to be conformationally unrestricted. For the models reported in Table 1, 67% were "good" according to the AlignScore metric; however of the 27 "bad" AlignScore models, 13 were pharmacophorically reasonable when compared to the gold standard (P15121, P27338, P46881, P00374, Patel_HIVRT, P49841, P06239, Q08499, P00811, O76074, P08709, P07900, P02766) by manual inspection and PIF similarity.

Third, the PIF similarity metric is also useful to gauge success in an objective and automatic manner. However, one limitation is that the score is a global one, which makes it difficult to interpret where the differences in two models are located. The individual PIF similarities for each probe are reported by FLAPpharm; however, these are also "global" for each probe type, i.e. a 0.67 similarity for the acceptor probe may indicate that two out of three acceptor regions are found or that all three regions are found but at a lower intensity than in the gold standard. We are currently investigating an approach that would cluster the PIF field points into regions, which would combine the benefits of the PIF similarity metric with that of the individual feature-based approaches.

CONCLUSIONS

Published pharmacophore elucidation approaches typically use a handful of data sets for validation. In this paper we are reporting a new benchmark data set for the validation of pharmacophore elucidation methods. The data set consists of 81 targets and 960 ligands, taken from high resolution X-ray crystallographic data. The ligands have been aligned using their cocrystallized protein targets, to provide the experimentally derived gold standard alignment set. We have also reported a methodology to objectively determine whether or not the molecular alignments produced by the pharmacophore elucidation approach are in agreement with this gold standard. This methodology consists of three aspects; first determining whether the approach has been successful in reproducing the experimental X-ray conformations, second by aligning the model to the gold standard reference and determining what proportion of molecules are aligned successfully by calculating the rmsd to their X-ray equivalents (AlignScore metric), and third by measuring the pharmacophoric interaction field (PIF) similarity between the model and the gold standard. This benchmarking approach has been applied using our pharmacophore elucidation program FLAPpharm. For the unbiased models, the bioactive conformations were identified by the models in 67% of the ligands on average, and 83% of the data set models contained ligands with a high number (>75%) of successfully predicted bioactive conformations.

Successful models according to the AlignScore metric were found for 67% of the data set, and the average PIF similarity of the best model to the gold standard was 74%. However, a significant number of the models that failed the AlignScore metric were found to be pharmacophorically reasonable; including these, the number of successful models rises to 83%.

Using the X-ray conformations as input to the method (rigid alignment), successful models were found for 93% of the data set, and the average PIF similarity of the best model to the gold standard was 89%. It is our view that this demonstrates not only that FLAPpharm is an extremely useful approach in pharmacophore elucidation, but also that this benchmarking data set and objective validation methodology is a valuable tool in the evaluation of pharmacophore elucidation methods. To this end we have made it available at http://www.moldiscovery.com/PharmBench along with a web service to score a user-provided alignment set to the gold standard using the metrics described above.

AUTHOR INFORMATION

Corresponding Author

*E-mail: simon@moldiscovery.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to acknowledge Fabrizio Buratta for helping develop and maintain the www.moldiscovery.com/PharmBench web page. F.M. acknowledges Commissione Europea, Fondo Sociale Europeo e della Regione Calabria, for their financial support.

REFERENCES

(1) 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**, DOI: 10.1021/ci300153d.
(2) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.
(3) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279−294.
(4) 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.
(5) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(6) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178.
(7) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50*, 1442−1450.
(8) Matter, H.; Will, D. W.; Nazaré, M.; Schreuder, H.; Laux, V.; Wehner, V. Structural Requirements for Factor Xa Inhibition by 3-Oxybenzamides with Neutral P1 Substituents: Combining X-ray Crystallography, 3D-QSAR, and Tailored Scoring Functions. *J. Med. Chem.* **2005**, *48*, 3290−3312.
(9) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.
(10) 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.
(11) Wolber, G.; Langer, T. LigandScout: 3D pharmacophores derived from protein bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.
(12) Feng, J.; Sanil, A.; Young, S. S. PharmID: pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **2006**, *46*, 1352−1359.
(13) 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.
(14) Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 773−788.
(15) *LigandScout*, version 3.0; Inteligand: Vienna, Austria, 2010; http://www.inteligand.com (accessed August 6, 2012).
(16) Ortuso, F.; Langer, T.; Alcaro, S. GBPM: GRID-based pharmacophore model: concept and application studies to protein-protein recognition. *Bioinformatics* **2006**, *22*, 1449−1455.
(17) Alcaro, S.; Artese, A.; Ceccherini-Silberstein, F.; Chiarella, V.; Dimonte, S.; Ortuso, F.; Perno, C. F. Computational analysis of Human Immunodeficiency Virus (HIV) type-1 reverse transcriptase crystallographic models based on significant conserved residues found in Highly Active Antiretroviral Therapy (HAART)-treated patients. *Curr. Med. Chem.* **2010**, *17*, 290−308.
(18) Leach, A. R.; Gillett, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539−558.
(19) Jones, G. GAPE: An Improved Genetic Algorithm for Pharmacophore Elucidation. *J. Chem. Inf. Model.* **2010**, *50*, 2001−2018.
(20) Drugport. http://www.ebi.ac.uk/thornton-srv/databases/drugport (accessed August 1, 2011).
(21) RCSB Protein Data Bank. http://www.rcsb.org (accessed August 5, 2011).
(22) RCSB Protein Data Bank—Ligand Expo. http://ligand-expo.rcsb.org (accessed August 5, 2011).
(23) Electron Density Server. http://eds.bmc.uu.se/eds (accessed August 5, 2011).
(24) (a) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *9*, 739−747. (b) *The PyMOL Molecular Graphics System*, Version 1.3; Schrödinger, LLC: New York, 2010.
(25) *SYBYL X*, version 1.3; Tripos, L.P.: St Louis, 2011.
(26) Gardiner, E. J.; Cosgrove, D. A.; Taylor, R.; Gillet, V. J. Multiobjective optimization of pharmacophore hypotheses: Bias toward low-energy conformations. *J. Chem. Inf. Model.* **2009**, *49*, 2761−2773.