JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

SERAPhiC: A Benchmark for in Silico Fragment-Based Drug Design

Angelo D. Favia,*[,†] Giovanni Bottegoni,[†] Irene Nobeli,[‡] Paola Bisignano,[†] and Andrea Cavalli*[,†,§]

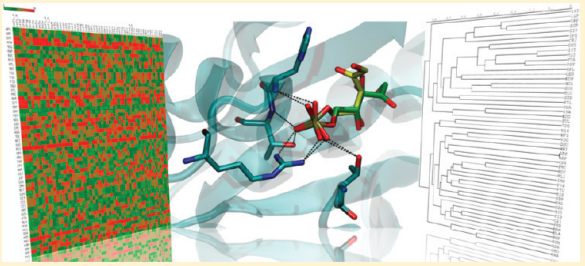
[†]Department of Drug Discovery and Development, Istituto Italiano di Tecnologia, via Morego 30, 16163 Genova, Italy

[‡]Department of Biological Sciences, Institute of Structural and Molecular Biology, Birkbeck, University of London, Malet Street, WC1E 7HX London, United Kingdom

[§]Dipartimento di Scienze Farmaceutiche, Università di Bologna, via Belmeloro 6, 40126 Bologna, Italy

Supporting Information

ABSTRACT: Our main objective was to compile a data set of high-quality protein−fragment complexes and make it publicly available. Once assembled, the data set was challenged using docking procedures to address the following questions: (i) Can molecular docking correctly reproduce the experimentally solved structures? (ii) How thorough must the sampling be to replicate the experimental data? (iii) Can commonly used scoring functions discriminate between the native pose and other energy minima? The data set, named SERAPhiC (**Se**lected **Fra**gment **P**rotein **C**omplexes), is publicly available in a ready-to-dock format (http://www.iit.it/en/drug-discovery-and-development/seraphic.html). It offers computational medicinal chemists a reliable test set for both in silico protocol assessment and software development.



INTRODUCTION

Fifteen years after their introduction,[1] fragment-based (FB) approaches are firmly established tools for drug discovery projects in which low molecular weight molecules (100−300 Da) are screened against biological targets of pharmaceutical interest.[2] Because of their low binding affinity for macromolecules (100 μM to 10 mM), fragments can only be assayed by highly sensitive biophysical techniques such as surface plasmon resonance spectroscopy (SPR),[3] NMR-based detections,[4,5] isothermal titration calorimetry (ITC),[6] and X-ray crystallography.[7] However, these experiments are generally time consuming and expensive. It is thus tempting to use computational techniques to prioritize molecules for wet assays, particularly during the earliest stages of the hit-identification process, thereby increasing the benefit/cost ratio. Molecular docking, for instance, is a routinely used drug discovery tool.[8] However, commonly used docking algorithms are typically tuned on data sets comprising bigger molecules with lead-like features. Consequently, despite some successful implementations of docking engines in fragment-oriented projects,[9−12] several issues are emerging around the transferability of the technique to this relatively new field.[13,14]

To explore these issues, Sándor and colleagues tested the docking accuracy of a well-known piece of software, Glide, on 190 protein−fragment complexes.[15] Despite the theoretical limitations, they reported some very encouraging outcomes. The software was assessed for its ability to reproduce experimentally obtained docking poses and to rank them adequately. Similarly, Kawatkar and co-workers tested various Glide scoring schemes on two test cases. By comparing the outcomes to experimental data, they showed that in silico predictions yielded considerably better results than random predictions.[16] Notably, an MM-GBSA rescoring procedure was used to improve the final outcomes but with poor results. This is further evidence that the success rate of physics-based rescoring procedures are strongly case dependent, an observation already made by some of us elsewhere.[17] Physics-based rescoring procedures are among several reported protocols that aim to improve the docking results of fragment screenings. For instance, Friedman and Caflisch applied a consensus scoring to high-throughput docking outcomes. This aided the discovery of 13 plasmepsin binders belonging to three different chemotypes.[18] Others have used QM methods to more accurately mimic polarization effects in the protein environment upon binding.[19] In this case, the methodology outperformed a prototypical scoring function, such as the one used in GOLD,[20] in identifying the correct binding modes. A few years ago, Marcou and Rognan tackled the issue from a different standpoint. Their method used the preacquired data of several experimentally obtained fragment−protein complexes. When reproducing known structural data, their method proved to be more accurate than four standard docking engines.[13] In their work, knowledge about the binding modes of several fragments to proteins was conveniently encoded in monodimensional vectors (IFPs). These were used to bias the conformational search toward more experimentally probable configurations.

There is growing interest in fragment docking, both in industry and in academia.[21] However, to the best of our knowledge, there is still no publicly available and purpose-built collection of

Received: July 21, 2011
Published: September 21, 2011

Table 1. X-ray Complexes in SERAPhiC

PDB id	protein	fragment (id)	substrate/ inhibitor	potency[a]	cofactor	metal coord.	res. (Å)	EC	CATH	source organism
1e2i	thymidine kinase	APR/APS	inhibitor	K_i = 5.3 mM			1.9	2.7.1.21	3.40.50.300	*Herpes simplex*
1f5f	sex-hormone binding globulin	DHT	substrate	IC_{50} = 3.8 nM[a]			1.7	NA	2.60.120.200	*Homo sapiens*
1f8e	neuramidase	49A	inhibitor	K_i = 15 μM			1.4	3.2.1.18	2.120.10.10	*Influenza virus*
1fsg	hypoxantine-guanine phosphoribosyltransferase	9DG	inhibitor	IC_{50} = 9.5 μM			1.05	2.4.2.8	3.40.50.2020	*Toxoplasma gondii RH*
1h46	exoglucanase	RNP	inhibitor	K_i = 270 μM			1.52	3.2.1.91	2.70.100.10	*Phanerochaete chrysosporium*
1k0e	*p*-aminobenzoate synthase	TRP	inhibitor				2	2.6.1.85	3.60.120.10	*Escherichia coli*
1m2x	metallo-beta-lactamase	MCO	inhibitor	K_i= 70−100 μM		Zn	1.5	3.5.2.6	3.60.15.10	*Elizabethkingia meningoseptica*
1m3u	ketopantoate transferase	KPL	product			Mg	1.8	2.1.2.11	3.20.20.60	*Escherichia coli*
1mlw	tryptophan 5-monooxygenase	HBI	inhibitor				1.71	1.14.16.4	1.10.800.10	*Homo sapiens*
1ofz	fucose specific lectin	FUL	inhibitor	K_d = 24.1 μM K_a = 4.15 × 10^4 M^{-1}			1.5	NA	2.120.10.70	*Aleuria aurantia*
1pwm	aldose reductase	FIS	inhibitor	IC_{50} = 9−35 nM	NAP		0.92	1.1.1.21	3.20.20.100	*Homo sapiens*
1r5y	queine trna-ribosyltransferase	DQU	inhibitor	K_i = 0.35 μM			1.2	2.4.2.29	3.20.20.105	*Zymomonas mobiliz*
1s5n	xylose isomerase	XYL	inhibitor				0.95	5.3.1.5	3.20.20.150	*Streptomices olivo-chromogenes*
1sd1	5′-methylthioadenosine phosphorylase	FMC	inhibitor				2.03	2.4.2.28	3.40.50.1580	*Homo sapiens*
1sqn	progesterone receptor	NDR	agonist	K_d = 0.4 nM			1.45	NA	1.10.565.10	*Homo sapiens*
1t0l	isocitrate dehydrogenase	ICT	substrate		NAP	Ca	2.41	1.1.1.42	3.40.718.10	*Homo sapiens*
1tku	3,4-dihydroxy-2-butanone 4-phosphate synthase	5RP	substrate				1.66	4.1.99.12	3.90.870.10	*Candida albicans SC5314*
1ui0	uracil-dna glycosylase	URA	inhibitor	K_i = 88 nM			1.5	3.2.2	3.40.470.10	*Thermus thermophilus HB8*
1uwc	feruloyl esterase	FER	product				1.08	3.1.1.73	3.40.50.1820	*Aspergillus niger*
1w1a	polysaccharide deacetylase	NDG	inhibitor				2.25	3.5.1.104	3.20.20.370	*Bacillus subtilis*
1wog	agmatinase	16D	inhibitor				1.8	3.5.3.11	3.40.800.10	*Deinococcus radiodurans*
1x07	undecaprenyl pyrophosphate synthetase	IPE	substrate	K_d = 520 μM		Mg	2.2	2.5.1.31	3.40.1180.10	*Escherichia coli*
1xfg	glucosamine-fructose-6-phosphate aminotransferase	HGA	inhibitor	K_i = 15 μM			1.85	2.6.1.16	3.60.20.10	*Escherichia coli*
1y2k	camp-specific 3′,5′-cyclic phosphodiesterase	7DE	inhibitor	IC_{50} = 21 nM			1.36	3.1.4.17	1.10.1300.10	*Homo sapiens*
1yki	oxygen insensitive nad(p)h nitroreductase	NFZ	substrate		FMN		1.7	1.5.1.34	3.40.109.10	*Escherichia coli*
1ynh	succynilarginine dihydrolase	SUO	product				1.95	3.5.3.23	3.75.10.20	*Escherichia coli*
1yv5	farnesyl pyrophosphate synthetase	RIS	inhibitor	IC_{50} = 5.7 0 K_i = 16 nM		Mg	2	2.5.1.1; 2.5.1.10	1.10.600.10	*Homo sapiens*
2aie	peptide deformylase	SB9	inhibitor	IC_{50} = 2.2		Ni	1.7	3.5.1.88	3.90.45.10	*Streptococcus pneumoniae*
2b0m	dihydroortate dehydrogenase	201	inhibitor	K_i = 2.8 μM	FMN		2	1.3.5.2	3.20.20.70	*Homo sapiens*
2bkx	glucosamine-6-phosphate deaminase	F6R	product				1.4	3.5.99.6	3.40.50.1360	*Bacillus subtilis*
2bl9	dihydrofolate reductase-thymidylate synthase	CP6	inhibitor	K_i = 0.16 nM	NDP		1.9	1.5.1.3; 2.1.1.45	3.40.430.10	*Plasmodium vivax*
2brt	leucoanthocyanidin dioxygenase	NAR	substrate				2.2	1.14.11.19	2.60.120.330	*Arabidopsis thaliana*
2cix	chloroperoxidase	CEJ	substrate	K_d = 33 mM	HEME		1.8	1.11.1.10	1.10.489.10	*Caldariomyces fumago*
2f6x	(*S*)-3-*o*-geranylgeranylglyceryl phosphate synthase	1GP	substrate				2	2.5.1.41	3.20.20.390	*Archaeoglobus fulgidus*
2fdv	cytochrome p450	D2G	inhibitor	K_i = 0.8 μM	HEME	Fe	1.65	1.14.14.1	1.10.630.10	*Homo sapiens*
2ff2	iag-nucleoside hydrolase	IMH	inhibitor	K_i = 6.2 nM		Ca	2.2	3.2.2.1	3.90.245.10	*Trypanosoma vivax*

dx.doi.org/10.1021/ci2003363 |*J. Chem. Inf. Model.* 2011, 51, 2882–2896

Table 1. Continued

PDB id	protein	fragment (id)	substrate/ inhibitor	potency[a]	cofactor	metal coord.	res. (Å)	EC	CATH	source organism
2fgq	outer membrane porin protein	MLT	substrate				1.45	NA	2.40.160.10	*Delfitia acidovorans*
2gg7	methionine aminopeptidase	U14	inhibitor	IC_{50} = 1.75 μM		Co	1.12	3.4.11.18	3.90.230.10	*Escherichia coli K12*
2gvv	phosphotriesterase	DI9	inhibitor	K_i = 125 μM		Ca	1.73	3.1.8.2	2.120.10.30	*Loligo vulgaris*
2hdq	beta-lactamase	C21	inhibitor	K_i = 40 mM			2.1	3.5.2.6	3.40.710.10	*Escherichia coli K12*
2i5x	receptor-type tyrosine-protein phosphatase beta	UA5	inhibitor				1.7	3.1.3.48	3.90.190.10	*Homo sapiens*
2iba	uricase	AZA	inhibitor				1.5	1.7.3.3	3.10.270.10	*Aspergillus flavus*
2j5s	beta-diketone hydrolase (hydrolase)	KTA	product				1.57	3.7.1.7	3.90.226.10	*Anabaena sp PCC 7120*
2p1o	skp1-like protein	NLA	substrate		IPH		1.9	NA	3.30.710.10	*Arabidopsis thaliana*
2q6m	cholix toxin	P34	inhibitor	K_d = 510 nM			1.25	2.4.2.30	3.90.175.10	*Vibrio cholerae*
2qwx	ribosyldihydronicotinamide dehydrogenase [quinone] (oxidoreductase)	ML1	inhibitor	K_i = 7.2 μM K_i = 92 μM	FAD		1.5	1.10.99.2	3.40.50.360	*Homo sapiens*
2rdr	1-deoxypentalenic acid 11-beta hydroxylase. fe(ii)/alpha-ketoglutarate dependent hydroxylase	OGA	inhibitor			Fe	1.7	1.14.15.4	2.60.120.620	*Streptomices avermitilis*
2uy5	endochitinase	H35	inhibitor	K_i = 3.2 μM			1.6	3.2.1.14	3.20.20.80	*Saccaromyces cerevisiae*
2v77	carboxypeptidase	PAY	inhibitor	K_i = 8.7 μM		Zn	1.6	3.4.17.1	3.40.630.10	*Homo sapiens*
2zvj	catechol *o*-methyltransferase	KOM	inhibitor	IC_{50} = 1.8 μM	SAM	Mg	2.3	2.1.1.6	3.40.50.150	*Rattus norvegicus*
3c0z	histone deacetylase	SHH	inhibitor	IC_{50} = 113 μM		Zn	2.1	3.5.1.98	3.40.800.20	*Homo sapiens*
3dsx	geranylgeranyl transferase type-2 subunit alpha	GER	product	K_d = 1.4 mM.			2.1	2.5.1.60	1.50.10.20	*Rattus norvegicus*
3eko	heat shock protein hsp 90-alpha	PYU	inhibitor	IC_{50} = 0.4 μM K_i = 0.2 μM C. IC_{50} > 20 μM			1.55	NA	3.30.565.10	*Homo sapiens*

[a] All values were extracted from the primary reference. The only exceptions are entries 1pwm and 1yv5, whose IC_{50} and K_i values, respectively, were taken from the BindingDB (http://www.bindingdb.org/).

dx.doi.org/10.1021/ci2003363 |J. Chem. Inf. Model. 2011, 51, 2882–2896

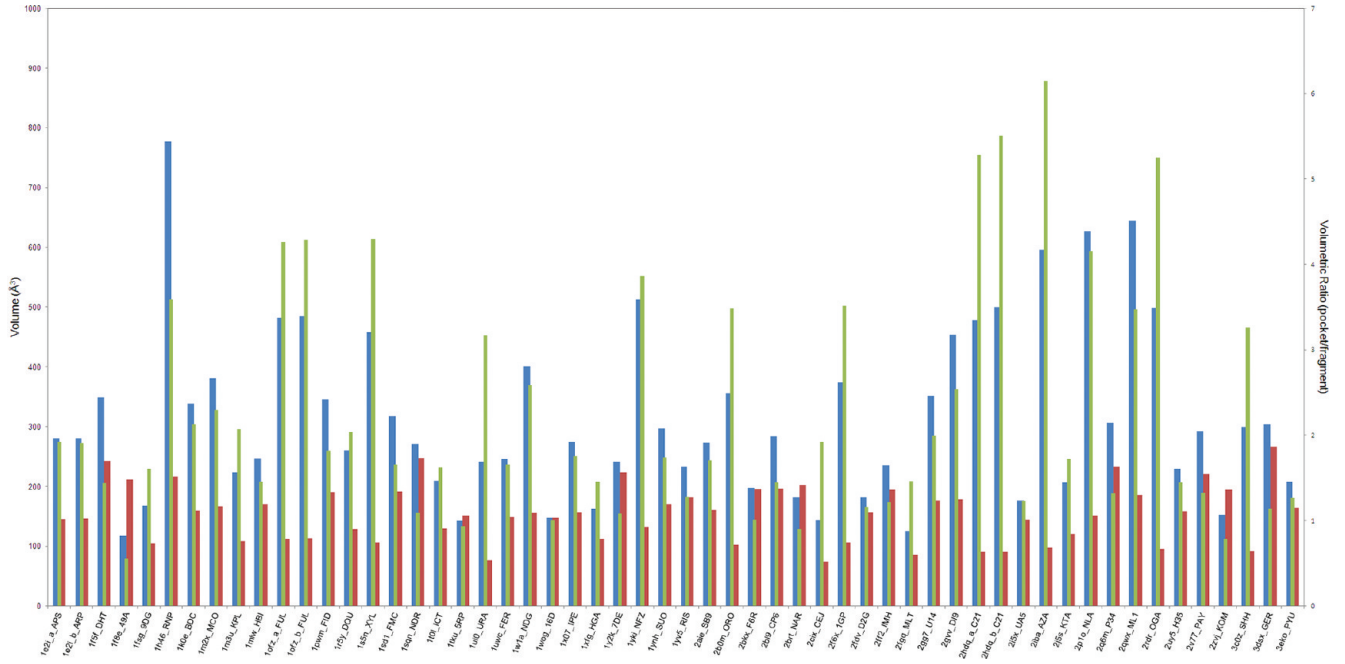


Figure 1. Analysis of the volumes of the pocket−fragment pairs included in SERAPhiC. Pocket and fragment volumes are highlighted in blue and red, respectively. The volumetric ratio between pockets and fragments is illustrated in green. The *y* axis is labeled according to volumes and volumetric ratio on the left and right, respectively.

reliable fragment−protein complexes for use as a validation tool. To address this gap, we assembled a set of high-quality protein−fragment complexes extracted from the Worldwide Protein Data Bank (wwPDB)[22] to enable the assessment of commonly used docking software and to assist the development of novel scoring functions, designed ad hoc to handle fragments. Each element of the wwPDB, as of June 2010, was filtered according to several criteria. This was to ensure the high quality of the final data set and include complexes of interest from a medicinal/biological standpoint. Every entry of the final data set had to meet the following conditions: (i) Resolution ≤ 2.5 Å, (ii) date of deposition ≥ year 2000, (iii) presence of an article describing the X-ray structure, (iv) macromolecule (only wild-type proteins, CATH[23] annotated, were considered) constituted by ≥ 200 a.a., and (v) presence of at least one biologically relevant small molecule bound (78 ≤ MW ≤ 300; heavy atoms count ≥ 6). The complexes that met these criteria were then clustered at the homologous superfamily level of CATH (only single-domain proteins were considered). This was to ensure the presence of diversified protein types. A single representative from each cluster was selected on the basis of source organism (human proteins were given priority) and resolution. These were eventually included in the definitive set. Finally, a manual curation allowed us to perform a careful check of the structures to incorporate only those proteins in complex with fragments that showed a density fit greater than 0.8 (see Methods). The filtering funnel was similar in spirit to the one used by Hartshorn and colleagues to derive the Astex diverse set.[24]

Finally, the data set was challenged by docking procedures to address the following questions: (i) Can molecular docking correctly reproduce the experimentally solved structures? (ii) How thorough must the sampling be to accurately replicate these structures? (iii) Can commonly used scoring functions discriminate between the true pose(s) and other energy minima? (iv) Can scoring functions adequately rank the native fragments with respect to decoys?

The data set, named SERAPhiC (**Se**lected **Fra**gment **P**rotein **C**omplexes), is publicly available in a ready-to-dock format (http://www.iit.it/en/drug-discovery-and-development/seraphic.html). It offers computational scientists a reliable test set for scoring-function assessment and code development.

RESULTS AND DISCUSSION

Data Set. SERAPhiC comprises 53 high-quality X-ray models of fragment−protein complexes, extracted from the wwPDB, according to the criteria detailed in the Methods section. All complexes are reported in Table 1 along with information such as the chemical name and function of the fragment, protein name, crystallographic resolution, and experimental in vitro or cellular data, where available.

Selection and Analysis of Complexes. Each entry of SERAPhiC has been the subject of a scientifically published study. This was necessary during the manual curation step to allow for an in-depth understanding of each structure's biological relevance. It also allows the interested user to delve into the articles for scientific or technical knowledge related to the deposited data. A number of X-ray models were discarded due to missing reference articles. This was expected in light of the trend in recent years for generating X-ray structures in a high-throughput manner.[25] Another important prerequisite was that the included entries had to have been deposited at the wwPDB on or after January 2000. This was to maximize the homogeneity of the database entry formats and focus attention on complexes likely to be of higher quality. Furthermore, we selected only entries where electron density maps had been deposited along with the structures. This allowed us to perform a quality check of the X-ray structures, especially at the fragment level (see below for further details), where uncertainty about the atom positions was not desirable.

dx.doi.org/10.1021/ci2003363 |*J. Chem. Inf. Model.* 2011, 51, 2882–2896

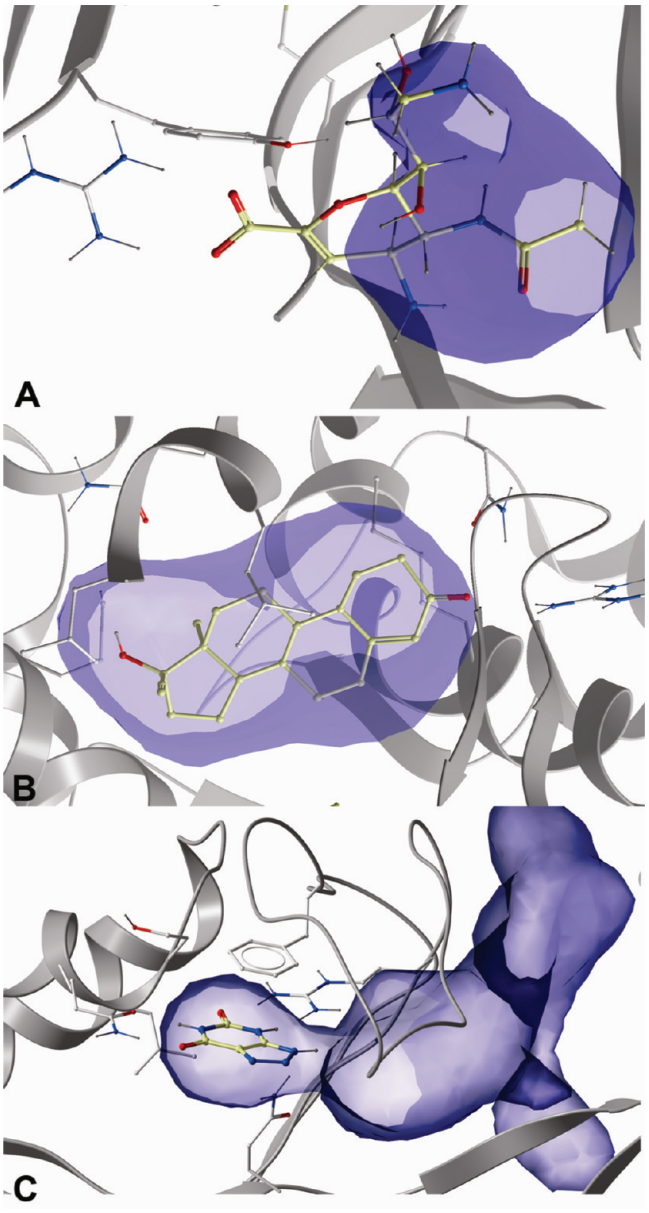


Figure 2. Three cases of varying ratios between pocket and fragment volumes. Entries 1f8e, 1sqn, and 2iba are shown in A, B, and C, respectively. The protein is shown as white ribbons. Relevant residues of the binding site are shown as stick models C-colored white. The fragment is C-colored yellow. The binding pocket is highlighted in transparent blue.

One of the main issues associated with the use of fragments in docking protocols is the unambiguous determination (more pronounced than lead-like molecules) of their binding mode. In fact, less stringent pharmacophoric constraints could result in alternative, equally good binding solutions. This makes it hard to neatly rank multiple, in-silico-generated poses. While this may seem to call for more fine-tuned scoring functions, it is worth noting that this behavior may also occur in nature more often than is usually acknowledged. The experiments of Allen and colleagues on multiple solvent crystal structures (MSCS) have shown how, by working in controlled excess of organic solvents, alternative configurations of small molecules at a target active site are possible.[26] With this in mind, the search for the fragment

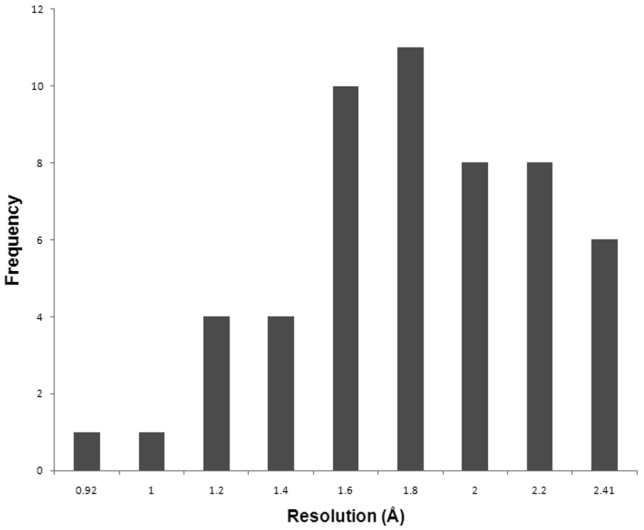


Figure 3. Histogram showing the distribution of the crystallographic resolutions over the entire SERAPhiC data set.

absolute binding configuration should perhaps be reconceptualized as the search for a set of possible multiple binding solutions. Interestingly, in support of this concept, some of the SERAPhiC entries include a protein in complex with a fragment binding to multiple sites (see the Analysis of Fragments section). For example, in entry 2hdq, the structure of β-lactamase was used to investigate whether fragments recapitulate the binding geometry recorded when included in larger molecules. In this entry, the small inhibitor carboxythiophene could be found at alternative binding sites.[27] Moreover, there is a correlation between multiple interaction spots and available volume at the target site. Indeed, we can easily hypothesize that the uncertainty associated with in-silico-predicted binding configurations is proportional to the size of the target pocket/cleft (see Docking Results).[24,28]

Figure 1 reports the volumes of fragments and pockets of all the entries included in the final data set together with their ratio.

The smallest volume for a fragment, as extracted from the X-ray structure, was recorded for uracil (76.26 Å^3), which was found cocrystallized with a DNA glycosylase from *Thermus thermophilus* and solved at a resolution of 1.50 Å (PDBid 1ui0). The largest volume for a fragment (303.3 Å^3) was seen for a geranyl derivative in complex with a transferase (PDBid 3dsx). As expected, a larger variety of pocket volumes were registered for proteins. The smallest volume for a pocket (117.9 Å^3) was found for the *Influenza virus* neuraminidase in complex with an inhibitor (PDBid 1f8e). Conversely, the largest volume for a pocket (777.4 Å^3) was observed for a cellobiohydrolase of *Phanerochaete chrysosporium* (PDBid 1h46). However, extreme care should be taken when analyzing these numbers separately for a given entry. A more relevant parameter, for a protein−fragment pair, is the ratio between the volume of the pocket and the volume of the fragment. This parameter is 1 when the small molecule is contained in the binding site and scarce additional volume is available. The ratio is lower than 1 when the fragment is partially contained by the protein and partially exposed to the solvent. It is higher than 1 when the compound lies in a bigger pocket with extra available space. An ideal benchmark for in silico fragment-based drug design should contain all the above-mentioned cases, which offer different challenges for computational screening. In SERAPhiC, several entries have a ratio

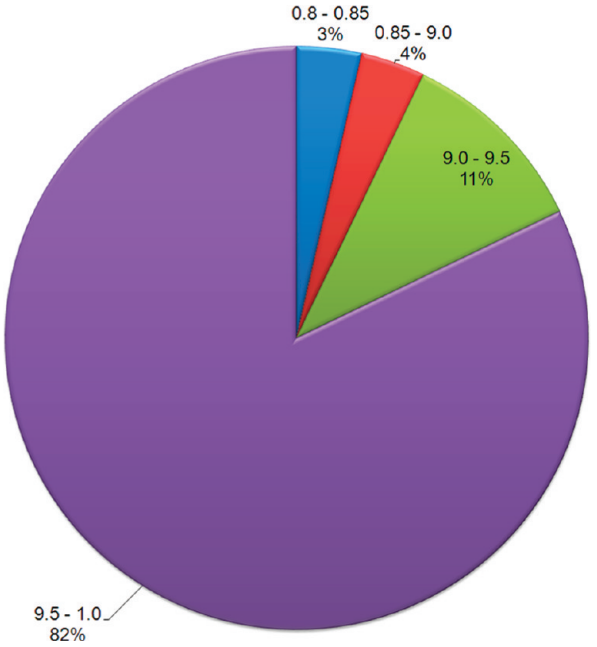


Figure 4. Pie chart summarizing the values of electron density fit recorded over the entire data set.

ranging between 1.4 and 2.2, i.e., the size of the enclosing pocket is markedly bigger than the size of the interacting fragment. The highest ratio (6.15) was for urate oxidase from *Aspergillus flavus* in complex with an inhibitor at 1.50 Å resolution (PDBid 2iba), while the lowest ratio (0.55) was for the previously mentioned complex between neuraminidase and 4,9-diamino-2-deoxy-2,3-dehydro-*N*-acetyl-neuraminic acid (PDBid 1f8e). In the latter case, the fragment was partially exposed to the solvent. One notable example is the complex between norethindrone and progesterone receptor (PDBid 1sqn), where the observed volumetric ratio was approximately 1. In Figure 2, these cases are illustrated to depict the distinct scenarios in the data set, which one might encounter while docking molecules at protein active sites.

By design, none of the fragments included in SERAPhiC were covalently bound to the interacting proteins. In a few cases, the interactions were mediated by metal ions (13 out of 53) or cofactors (9 out of 53). Mostly, however, direct interactions with amino acid side chains were present. Furthermore, we recorded four distinct entries (PDBids 3dsx, 2qwx, 2p1o, and 1yki), where the complexed fragment was located at the interface between two chains.

The structural resolution was an important filtering criterion during our data set construction. In order to be included, a complex had to have a resolution equal to or better than 2.5 Å. As shown in Figure 3, the distribution of the resolutions of the final data set ranged from 0.92 to 2.41 Å and was centered on an exceptionally good mean value of 1.69 Å (median = 1.7 Å). Nominal resolution alone is not usually sufficient to attest to the structure quality. Here, however, when considered together with the other benchmark selection parameters, it does provide a practical indicator of the crystal structure reliability.[29]

Analysis of Proteins. All entries contained wild-type proteins with a minimum length of 200 amino acids. This was to avoid including very small proteins or small crystallized chunks of bigger sequences. Moreover, since the data set was medicine/biology driven, mutants were not included. As explained in the Methods section, the final entries of SERAPhiC were chosen as representatives of clusters at the homologous superfamily level of CATH.[23] Hence, by design, the chosen proteins were both structurally and evolutionarily diverse. The maximum global pairwise sequence identity between any two of the chains in our data set was 20.8% (1xfg:A vs 1ui0:A). Two plots showing the distribution of all pairwise sequence identities are available as Supporting Information.

When selecting a cluster, representative human proteins were chosen first, where available, and then those with the highest resolution were given priority. Human proteins were privileged because we anticipated that end users would mostly be interested in pharmacological problems concerning humans. Moreover, although we recognize the usefulness and validity of NMR structures, we decided to include only models coming from X-ray studies. This was to avoid the extra complexity of including multiple conformations of the biomolecular complexes and to keep the data set as homogeneous as possible.

Analysis of Fragments. Part of SERAPhiC's structure is that each entry contains a fragment molecule that establishes noncovalent interactions with a protein. Hence, it was important to be confident about the fragments' positions. An electron density fit value was calculated for each small molecule in the data set, as detailed in the Methods section. This value indicates how well the observed electron density correlates with the position of the atoms in the fragment structure. It can assume values between −1 and 1, where the former means no fit at all and the latter means a perfect fit between the fragment and the electron density map. As highlighted in Figure 4, the vast majority of fragments (46 out of 56, i.e., 82%) had an exceptionally good electron density fit, with values between 0.95 and 1.0, while only 18% had a density fit between 0.8 and 0.95.

Frequent binders such as crystallization additives did not qualify for our purposes as biochemically interesting molecules. This assessment has already been proposed by others.[24,30,31] Proteins complexed with these chemicals only were filtered out from the data set. In addition to the list compiled by Hartshorn and colleagues, we did not consider tartaric acid (TLA) as a fragment. TLA occurred 70 times in the initial wwPDB filtering stages. The final list of excluded frequent binders is available as Supporting Information.

Some proteins were found in complex with a fragment binding at diverse sites (i.e., PDBids 1ofz and 2hdq). In the previously mentioned entry 2hdq, carboxythiophene molecules were found in multiple binding spots. We referred to the original publication in order to consider the two poses most likely to be pharmaceutically relevant. Both showed very good density fits (>0.94).[27] Entry 1ofz was a crystal model of lectin, a fungal protein in complex with both anomers of fucose. In particular, the β anomer of the sugar (i.e., the one included in the data set) was found bound to two very similar, although distinct, binding sites of the six-bladed β-propeller fold.[32] Finally, a case worth reporting is entry 1e2i, where the final density distribution did not allow a clear distinction between the two enantiomers of the inhibitor 9-hydroxypropyladenine, with both binding in two alternative ways.[33] For this reason, in the deposited model the authors included the (*R*) and (*S*) forms of the fragment, with both showing good density fits (0.82 and 0.91, respectively). These were particularly intriguing cases to process via molecular docking, because both of the possible binding poses of a single fragment have been experimentally observed and interaction with distinct molecules at the same protein site has been reported.

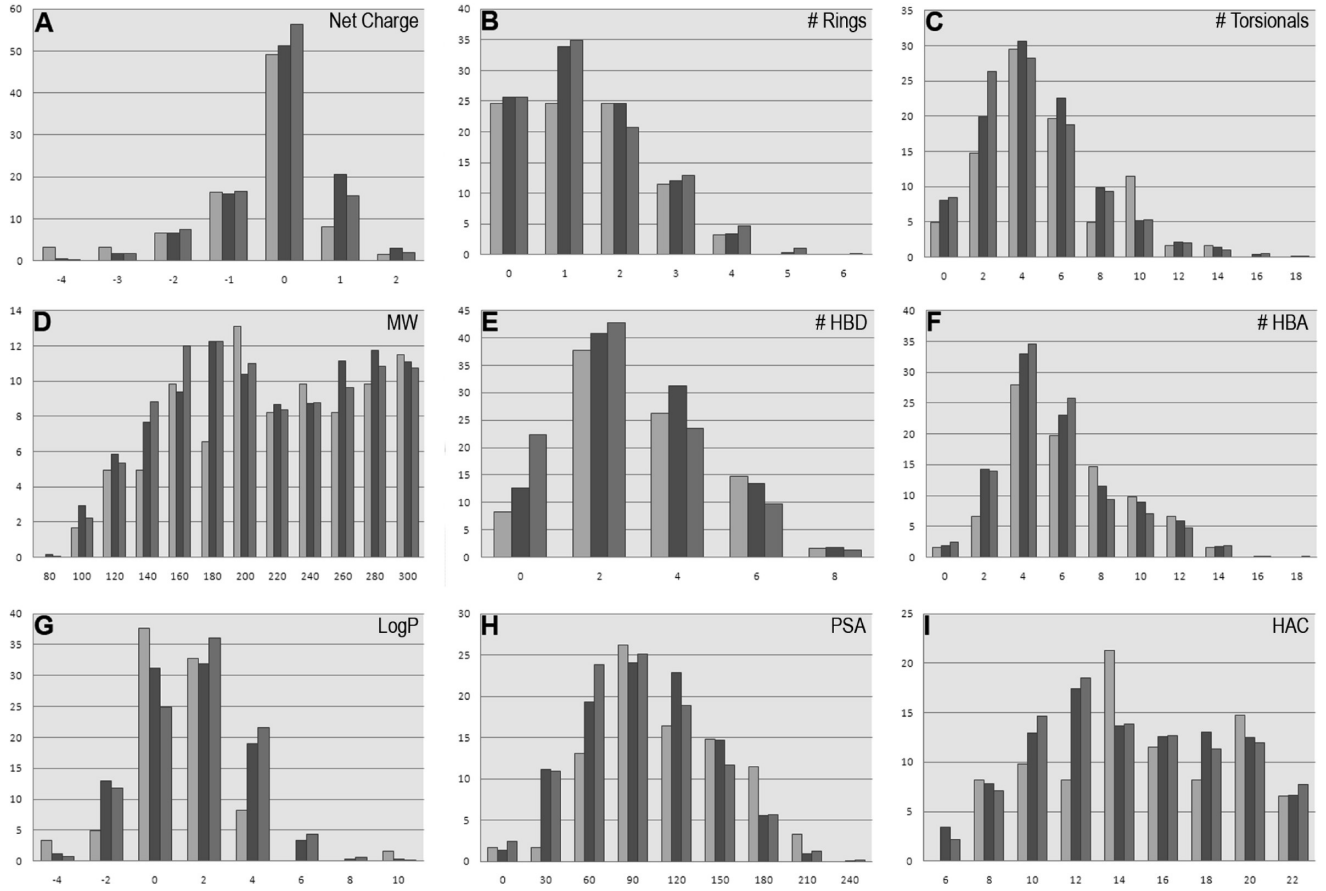


Figure 5. Percentage distribution of 9 molecular properties calculated for the small molecules included in SERAPhiC (shown in light gray), DrugBank (shown in dark gray), and KEGG COMPOUND (shown in gray). The selected properties are as follows: (A) Net charge, (B) number of rings, (C) number of torsionals, (D) molecular weight in Da, (E) HB donor, (F) HB acceptor, (G) log $P_{o/w}$, (H) polar surface area in Å^2, and (I) heavy atoms count.

We decided to include entries containing fragments that modulate the protein activity. These include both inhibitors and naturally occurring molecules such as substrates or products. We included the latter set of compounds because, in the past decade, docking has increasingly been used to search for naturally occurring ligands within pools of possible candidates for deorphanization purposes.[17,34−39] Moreover, natural modulators represent an extremely challenging task for molecular-mechanics-based virtual screenings. This is because, despite their naturally optimized complementarity for cognate biological targets, their binding is usually characterized by relatively weak affinity constants.[17] We also decided to include complexes where affinity or inhibition data is available and those without documented biochemical characterization. Nonetheless, a very high percentage (59%) of the entries in SERAPhiC was accompanied by some kind of biochemical assessment such as IC_{50}, K_i, or K_d determinations (see Table 1).

As mentioned, the presence of fragments bound was our most important criterion when parsing the wwPDB. With this in mind, one question is whether the extracted fragments are truly representative of bigger available sets of chemicals such as commonly used drugs or naturally occurring ligands of a similar size. If so, this would legitimize SERAPhiC as a fair benchmark for docking protocols and scoring function development in the field of fragment-based discovery. To assess this, we considered nine commonly used physicochemical molecular descriptors (i.e., net charge, no. of rings, no. of rotatable bonds, molecular weight, no. of HB donor, no. of HB acceptor, log $P_{o/w}$, polar surface area, and no. of heavy atoms). We compared their distributions over the entire fragment set to the distributions observed in similar-sized molecules in two publicly available data sets, namely, DrugBank[40] and KEGG COMPOUND.[41] The DrugBank database contains annotated drug entries, while the KEGG COMPOUND database is a collection of small molecules, biopolymers, and other chemical substances that are relevant to biological systems. The distributions for all three data sets are reported in Figure 5; each panel contains all three distributions for a given descriptor.

A comparative analysis shows significant similarities within the distributions of every single property of the three data sets. We note that a limited number of fragments were selected, and chemotype diversity was not a compilation criterion. Nonetheless, for SERAPhiC, normal distributions can be observed for most of the descriptors. Conversely, the highly correlated molecular weights and the heavy atoms count descriptors (panels D and I in Figure 5) produced non-Gaussian distributions with the highest peaks centered at 160 and 280 Da and 14 and 20 atoms, respectively. Interestingly, very similar behaviors were also registered in DrugBank and KEGG COMPOUND. There is thus a significant correlation in terms of the physicochemical description between SERAPhiC and both DrugBank or KEGG COMPOUND. This speaks to the potential of safely using SERAPhiC without fear of missing the important molecular properties that

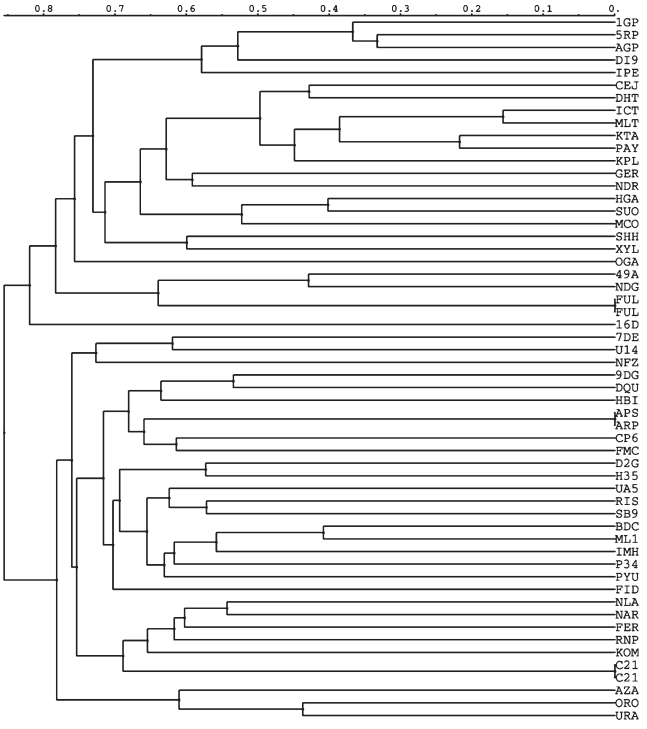


Figure 6. Clustering of the 56 fragments included in SERAPhiC using a fingerprint-based Tanimoto similarity matrix. Each fragment is labeled according to the three-letter code found in the corresponding PDB file.

are statistically found in drug-like molecules and naturally occurring ligands.

Nevertheless, physicochemical diversity does not directly account for the presence of sufficiently different chemotypes. To this end, we extended the fragments' characterization by clustering the entries according to their fingerprint similarity. In Figure 6, the 56 fragments of SERAPhiC are clustered according to their calculated Tanimoto pairwise distances between fingerprints (see Methods for details).

Fingerprinting is a useful way to capture the chemical information of a molecule and encode it into a monodimensional vector. It is particularly suitable for comparative purposes.[42] The readout of the cluster analysis (Figure 6) clearly shows how distinct the fragments are from each other. Clustering at a Tanimoto distance of 0.3 can be used to avoid the presence of overly similar chemotypes and to enrich the topological dissimilarity in the cluster representatives.[43] A 0.3 Tanimoto distance clustering of our data set produced 51 clusters. The final selection step of SERAPhiC was protein based (see Methods for further details). The initially unsought chemotype richness of the final set is thus a remarkable but welcome finding, which allows us to test computational protocols with fragments that cover a notable portion of the available chemical space.

Finally, we analyzed the SERAPhiC entries using BINANA,[44] the recently reported algorithm for ligand-binding characterization developed by Durrant and McCammon. Table 2 summarizes the interactions present for each fragment−protein complex in the final data set.

The most represented interactions in the data set were hydrophobic contacts (∼17 on average per fragment−protein pair), followed by H bonds (∼4.3 on average per pair), salt bridges (∼1.4 on average per pair), $\pi-\pi$ stackings (∼0.7 on average per pair), metal coordinations (∼0.5 on average per pair), T-stacking (∼0.3 on average per pair), and cation$-\pi$ interactions (∼0.07 on average per pair). The readout of this analysis illustrates the diversity of the interactions included in the set. Interestingly, the predominant presence of nonspecific hydrophobic interactions and the underrepresentation of interactions such as the T-stacking were not biased by the filtering criteria used at the entry selection step. In fact, very similar trends were observed in the data set of 2673 protein−ligand pairs used to validate BINANA. In the BINANA assessment, the relative ratios between types of interactions were mostly maintained. However, there was no upper constraint to the molecular weight of the complexed molecules (mean MW = 350 Da; SD = 172 Da). This resulted in an increased net number of interactions per pair. For instance, more than 5 H bonds and 21 hydrophobic contacts per protein−ligand pair were registered in the BINANA validation data set. The most significant difference between the BINANA validation set and SERAPhiC was detected for the cation$-\pi$ interactions, which were, on average, ∼0.33 per pair for the former with respect to the average value of ∼0.07 for the latter. This divergence was most likely because positively charged molecules were underrepresented in our data set (see Figure 5A).

Docking Results. It is beyond the scope of this work to test and compare the performances of commonly used docking programs when challenged with fragments. Our focus is on presenting and discussing this new set. Hence, we decided to use ICM3.7,[45] a well-established docking engine, to draw some preliminary conclusions about the use of fragments in a prototypical trial of computational docking. We exploited the data set with self-docking and a pilot-VLS procedure. In a self-docking exercise, the fragment is extracted from its active site. The ability of the software in finding the native pose is then measured. In a VLS-like trial, in contrast, every fragment is docked to every protein in a virtual-screening-like scenario. The fitness of the scoring function in matching up experimentally known pairs is then evaluated. Because we lacked experimental evidence, we followed a similar strategy as others[43] in assuming bona fide that (i) only configurations resembling the native one would occur biologically and (ii), if bioassayed, each protein would preferably interact with its native ligand rather than showing binding promiscuity for any other fragment in the data set.

The results of the self-docking simulations are reported in Figure 7, where the success rate over the entire data set, distinguished in soft and hard (top and bottom part of the plot, respectively), is plotted as a function of the Monte Carlo thoroughness.

'Soft success' is defined as the ability of the docking engine to find the native pose of the fragment within the top 10 scoring hits. 'Hard success' is concerned with the capacity of the docking algorithm to find the native pose of the molecule as an absolute top scoring configuration. In both cases, in line with other reported works,[15,16] we used a tolerance of 2 Å rmsd from the native pose. Moreover, three different incremental values of search volume were tested (within 3.5, 5, and 7.5 Å from the cognate fragment) to assess the relationship between success and dimensionality of the searchable space.

Hard success should always be the goal while benchmarking or tuning a docking procedure. However, soft success is a fair criterion for benchmarking novel protocols. In fact, from a drug discovery perspective, unless very large databases are screened, the top scoring pose is rarely the only one considered and trusted to be the experimentally occurring one. Hence, it is interesting to assess how many configurations one should produce in order to

Table 2. Count of Several Different Types of Interactions (as detected by BINANA) for the 56 Protein−Fragment Pairs Included in the Final Data Set

PDBid_chain_fragment	H bonds	hydrophobic contacts (C−C)	$\pi-\pi$ stacking	T-stacking	cation−π	salt bridges	metal coord.
1e2i_A_ARP	2	22	2	0	0	0	0
1e2i_A_APS	2	17	2	0	0	0	0
1f5f_A_DHT	2	20	0	0	0	0	0
1f8e_A_49A	9	10	0	0	0	6	0
1 fsg_A_9DG	5	34	3	2	0	0	0
1h46_X_RNP	1	25	3	0	2	3	0
1k0e_A_TRP	4	18	0	2	0	0	0
1m2x_A_MCO	3	13	0	0	0	2	2
1m3u_A_KPL	4	10	0	0	0	1	2
1mlw_A_HBI	1	26	4	0	0	0	0
1ofz_A_FUL_a	5	12	0	0	0	0	0
1ofz_A_FUL_b	9	48	0	0	0	0	0
1pwm_A_FID	5	18	1	0	0	0	0
1r5y_A_DQU	7	28	2	0	0	0	0
1s5n_A_XYL	5	17	0	0	0	0	0
1sd1_A_FMC	6	15	0	1	0	0	0
1sqn_A_NDR	0	23	0	0	0	0	0
1t0l_A_ICT	8	4	0	0	0	5	2
1tku_A_5RP	6	7	0	0	0	2	0
1ui0_A_URA	0	6	0	0	0	0	0
1uwc_A_FER	0	11	0	0	0	0	0
1w1a_1_NDG	6	8	0	0	0	0	0
1wog_A_16D	2	14	0	0	0	6	0
1 × 07_A_IPE	8	4	0	0	0	7	2
1xfg_A_HGA	12	2	0	0	0	3	0
1y2k_A_7DE	1	22	1	1	0	0	0
1yki_A_NFZ	3	12	0	0	0	0	0
1ynh_A_SUO	12	17	0	0	0	4	0
1yv5_A_RIS	5	8	0	0	0	4	5
2aie_P_SB9	5	18	0	0	0	0	2
2b0m_A_ORO	10	16	3	0	0	1	0
2bkx_A_F6R	11	5	0	0	0	2	0
2bl9_A_CP6	3	22	1	1	0	0	0
2brt_A_NAR	2	37	1	0	0	0	0
2cix_A_CEJ	0	23	0	3	0	0	0
2f6x_A_1GP	9	4	0	0	0	0	0
2fdv_A_D2G	2	32	0	2	0	0	1
2ff2_A_IMH	7	39	6	1	1	2	2
2fgq_X_MLT	6	4	0	0	0	4	0
2gg7_A_U14	1	16	0	2	0	2	2
2gvv_A_DI9	4	7	0	0	0	0	1
2hdq_A_C21_a	1	15	0	0	0	1	0
2hdq_A_C21_d	1	4	0	0	0	1	0
2i5x_A_UA5	9	18	0	0	1	2	0
2iba_A_AZA	7	9	3	0	0	0	0
2j5s_A_KTA	1	9	0	0	0	2	0
2p1o_B_NLA	3	25	0	0	0	3	0
2q6m_A_P34	2	41	4	0	0	1	0
2qwx_A_ML1	0	14	0	0	0	0	0
2rdr_A_OGA	4	7	0	0	0	4	2
2uy5_A_H35	4	42	5	1	0	0	0
2v77_A_PAY	11	10	0	0	0	7	2
2zvj_A_KOM	4	22	0	0	0	0	2

dx.doi.org/10.1021/ci2003363 |J. Chem. Inf. Model. 2011, 51, 2882–2896

Table 2. Continued

PDBid_chain_fragment	H bonds	hydrophobic contacts (C−C)	$\pi-\pi$ stacking	T-stacking	cation−π	salt bridges	metal coord.
3c0z_A_SHH	0	9	0	0	0	2	1
3dsx_A_GER	0	22	0	0	0	0	0
3eko_A_PYU	3	14	0	0	0	0	0

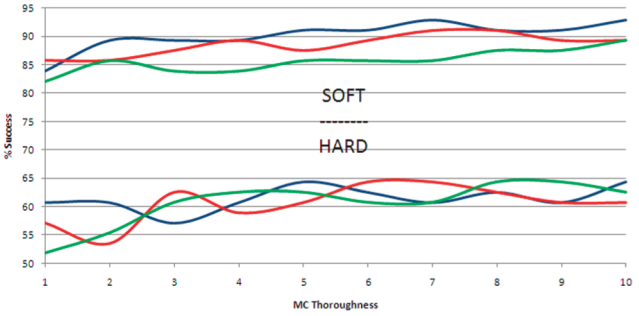


Figure 7. Soft (top) and hard (bottom) docking success percentages plotted as a function of the Monte Carlo thoroughness. The docking attempts were performed with different volume size grids, namely, 3.5, 5, and 7.5 Å from the cognate ligand (shown in blue, red, and green respectively).

increase the chances of including a pose that closely resembles the native one. As expected, enhancing the sampling generally produced better results until a limit was reached, which could be assessed as approximately 90% for soft success. Such a correlation was less straightforward for hard success. This was likely due to the strict criterion (top scorers only were considered) used to distinguish success from failure. Nonetheless, a 60% success rate was recorded on average. Finally, there were a few complexes that were not reproduced within the top 10 hits (PDBid 1tku, 2fgq, 2hdq, and 1qwx).

In the case of 1tku, the network of H-bond interactions of the phosphate group of the ribulose fragment with T85, R142, H145, and T146, present in the X-ray model, was exactly reproduced by the top scoring pose (see Figure 8A). The exceedingly high rmsd value of this pose was mostly caused by the differences found in the orientation of the other weakly interacting parts of the fragment. Rather than being an intrinsic limitation of docking,[46] this speaks to the inadequacy of a well-established metric (i.e., rmsd) to distinguish failure from success. Following the example of the recently reported work of Verdonk and colleagues,[47] we also used 1.5 Å as rmsd tolerance to analyze our data. However, no significant differences were recorded. An interaction-based metric, such as the already mentioned IFPs[13] or the structural interaction fingerprints (SIFt),[48,49] would likely have been more able to label the simulation as successful. In support of this, the 1tku protein rewarded its cognate fragment as top scorer among a crowd of possible alternative candidates in the virtual screening simulation (see below). This speaks to the strength and high specificity of the established interactions. The same considerations could be applied to entry 2fgq, where the docked fragment malate reenacted the same driving interactions found in the original X-ray model (i.e., H bonds with T36, R75, and R133). However, a different allocation of the carboxylic tail of the fragment resulted in formation of an alternative H bond with the side chain of T102. This produced a high value of rmsd (see Figure 8B). A comment is required for one of the two carboxythiophene fragments (PDBid 2hdq), where the original partial $\pi-\pi$ stacking with Y221 was kept in the fifth top scoring pose while an alternative orientation of the acidic group of the fragment, which yielded an additional H bond, was found within the large β-lactamase binding site (see Figure 8C). Ultimately, in the case of entry 2qwx, the top scoring pose of the indole derivative found in our docking simulations reproduced the original $\pi-\pi$ stacking with the FAD cofactor. Additionally, this pose resulted in a supplementary H bond with N161 that was not present in the starting X-ray structure (see Figure 8D). Interestingly, the top-ranked pose had a comparable if not slightly higher density fit value with respect to the deposited model (0.897 vs 0.887).

While the self-docking experiments deal with the ability of docking software to reproduce an expected binding mode, the pilot-VLS trials address the sensitivity of scoring functions in ranking the true binder either as top hit or among the best scorers when included in a set of diverse nonbinding molecules. This is similar to the standard structure-based screening of a fragment library, where only the topmost fraction of binders would be considered for experimental testing. The results of our VLS-like protocol are summarized in Figure 9 as a heat map, where each row represents a fragment and each column a protein, and each pixel of the matrix is colored according to the rank achieved in the simulation.

In an ideal scenario, each protein should reward the cognate fragment, the other molecules present should function as decoys, and the matrix should be green on the diagonal where the matching pairs occur. Although our data set was limited, an encouraging trend was observed. The cognate ligand was ranked 15 times as absolute top binder and ranked within the top five binders 24 times (out of 56 attempts). Fragments are more likely to show promiscuous binding, both in vitro and in silico. Considering the constitutive limitations of the scoring function used, this is thus an interesting result. Moreover, in the case of small-sized molecules, the effect of induced fit in proteins should play a minor role compared to simulations involving bigger compounds.[21] The occurrence of rewards triggered by a structure-induced bias should thus be limited to just a few examples, if any. Notably, there were some fragments that showed surprising complementarity for their natural counterparts. A few fragments also achieved good scores in most of the cases. The most promiscuous fragments were uracil (URA), deazaguanine (9DG), 8-azaxanthine (AZA), orotate (ORO), and carboxythiophene (C21), which always achieved very competitive scores regardless of the protein investigated. In contrast, fragments like 4,9-amino-2,4-deoxy-2,3-dehydro-*n*-acetyl-neuraminic acid (49A), dihydrotestosterone (DHT), norethindrone (NDR), risedronate (RIS), and octane-1,3,5,7-tetracarboxylic acid (PAY) showed a notable preference for their natural counterpart. To investigate the reason for the promiscuous behavior, we performed a principal component analysis (PCA) for the above-mentioned fragments. This was based on nine physicochemical descriptors (i.e., net charge, no. of rings, no. of rotatable bonds, molecular weight, no. of HB donor, no. of HB acceptor, log $P_{o/w}$, polar surface area,

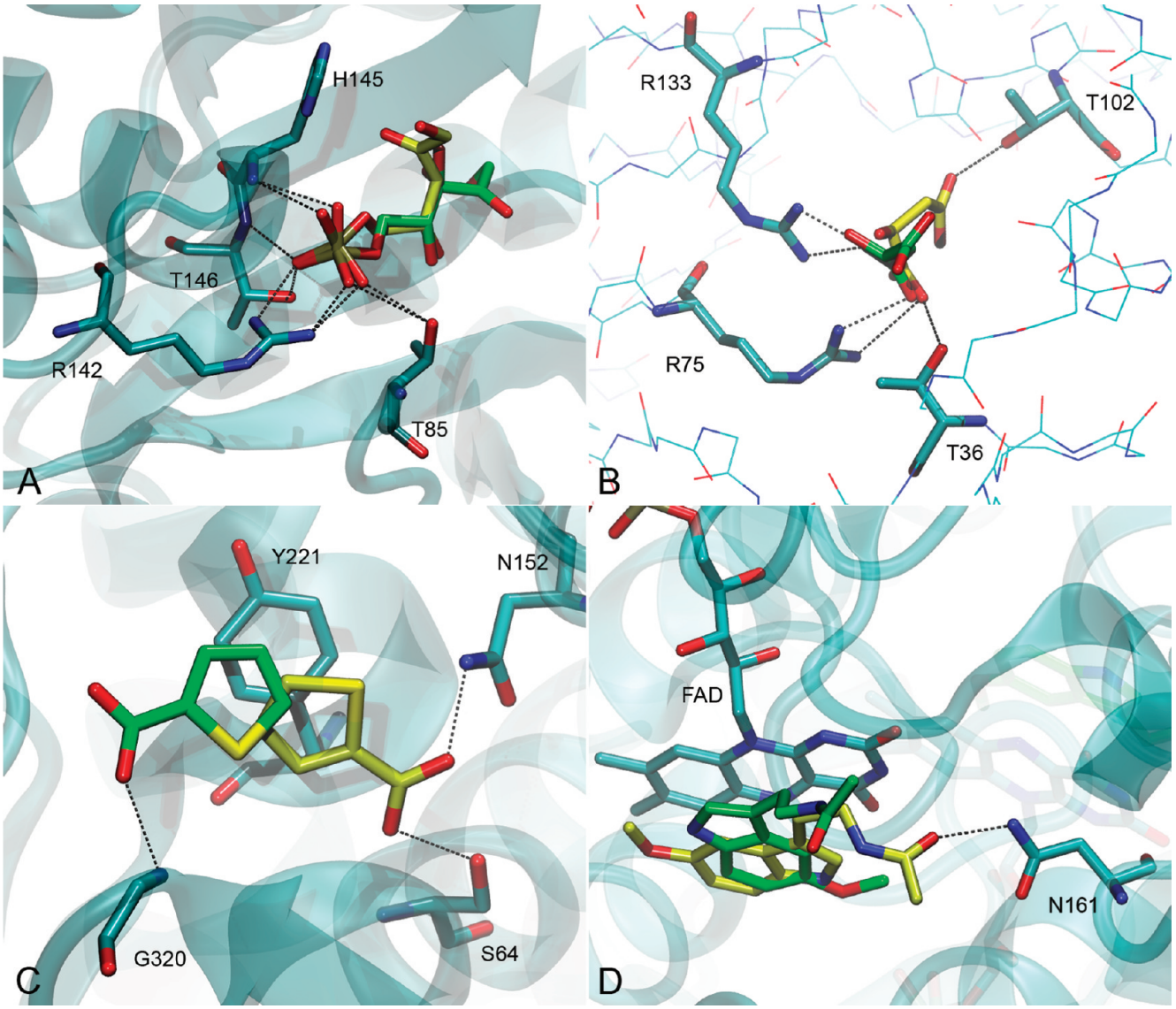


Figure 8. Overlay between selected fragments, as complexed in the X-ray models, and some corresponding top scoring poses, as obtained in the docking runs. Entries 1tku, 2fgq, 2hdq, and 2qwx are illustrated in panels A, B, C, and D, respectively. Proteins and important aminoacids are highlighted in cyan. The native pose is highlighted C-colored in green, while the docked pose is C-colored in yellow.

and no. of heavy atoms). The first two PCs explained 99% of the total variance. In particular, the first PC was sufficient to detach promiscuous fragments from selective ones. The descriptors that most contributed to the first PC were the highly correlated molecular weight and heavy atoms count and the polar surface area. Promiscuous fragments were, on average, smaller than selective molecules (139 vs 288.6 Da). They had a smaller polar surface area too (62 vs 84 Å^2). A plot illustrating this trend is available as Supporting Information. Another consideration is that, in a VLS-like study, the nature of the fragment can have an effect on the docking ranks that is even more relevant than the physicochemical complementarity with the studied target. Taken together, these data are in line with the hypothesis proposed by Hann and colleagues in 2001 and later confirmed experimentally by others that smaller molecules have a higher probability of hitting biological targets and are thus suitable starting points for drug discovery campaigns.[50,51]

We analyzed the docking ranks of the pilot-VLS using ligand efficiency (LE)[52] as the scoring function. This did not highlight any significant differences with respect to the results obtained with the standard ICM scoring function. A heat map showing the LE analysis is available as Supporting Information.

To assess the importance of the solvent in mediating the interactions between fragment and protein, all docking simulations were repeated, with the receptor definition for the map generation now including water molecules that bridged fragments to the proteins. In contrast to other studies,[24] the docking figures did not change significantly (data not shown).

Finally, we analyzed the relationship between success rates and the size of the pockets. This was simple for self-docking because the docking dimensionality increases with the search space. With the VLS-like framework, the availability of possible alternative hot spots in bigger sites could increase the chance that ligands would establish good interactions, even with noncognate proteins, thus achieving competitive scores.

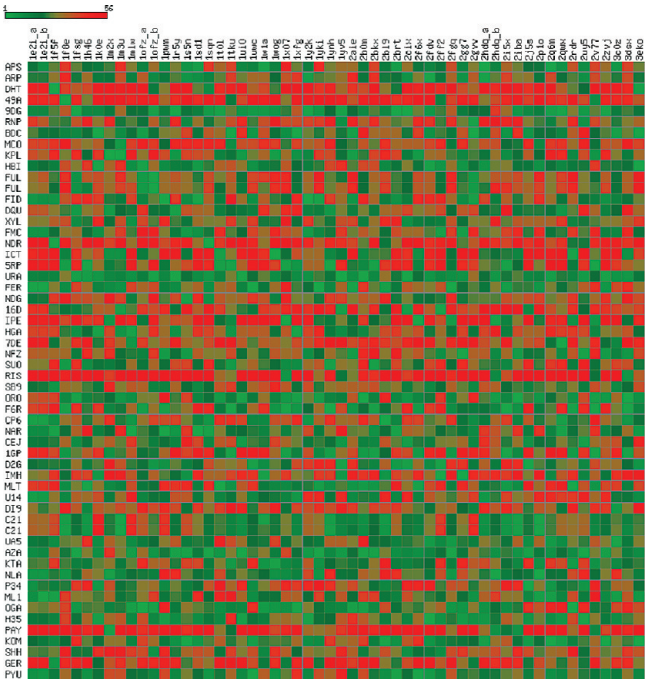


Figure 9. Heat map summarizing the pilot-VLS docking outcomes. Each row represents a fragment (labeled according to its three-letter code), and each column represents a protein (labeled according to the corresponding PDBid). Each pixel is colored according to the rank achieved during the docking simulations. Top binders are highlighted in green, while the poorest ranked fragments are depicted in red. The intermediate positions are colored as blends of varying ratios of green and red calculated according to the ranks achieved.

CONCLUSIONS

The main raison d'être of SERAPhiC was the need for a reliable and publicly available data set of protein−fragment complexes. Given the growing interest in fragment-based approaches, both in industry and in academia, we believe the time is right to use ad hoc computational tools to complement experiments in fragment-based drug discovery pipelines. Unfortunately, the prompt applicability of standard in silico techniques, such as molecular docking, to fragment-related matters seems to be greatly hampered by their constitutive inadequacy in dealing with small-sized molecules.[13,14] In this context, SERAPhiC was compiled to be a benchmark for assessing existing docking protocols and available scoring functions and for developing novel computational tools. The analysis performed on the fragments of the data set showed a good coverage of the physicochemical features statistically found in similar-sized molecules.

Ultimately, although it was not one of the filters applied during compilation, most of the entries of SERAPhiC were characterized by experimentally obtained biological data, such as K_i, IC_{50}, or K_d. These may be conveniently used as qualitative indicators when developing novel methods for predicting the free energy of association. Moreover, the diversity of fragments and proteins included in the data set ensures the presence of a wide variety of chemical interactions, each presenting different challenges for docking algorithms. Hence, the use of SERAPhiC as a validation tool is likely to guarantee a comprehensive coverage of the typical pharmacophoric features that can be found in a drug design effort.

The data set was challenged by self-docking and VLS-like procedures with a prototypical docking software to enact a standard simulative protocol. The analysis of the docking results showed that, in most cases, it was possible to recapitulate the experimental pose of the fragment within the cognate protein's binding site. The success rate of the self-docking procedures was slightly above 60% when the absolute top-ranked pose was the only solution considered. This rose to approximately 93% when the top 10 scoring configurations were taken into account. In four cases, the native pose was not found within the top 10 configurations and an extended search was necessary. Nonetheless, it was shown that, in such cases, an interaction-based metric would have produced better results.

Finally, we performed a pilot-VLS docking study. Although the data set used was limited in size, some important conclusions were possible. The effects of binding promiscuity were registered for several fragments. Fragment multivalency was mostly associated with lower molecular weights, where the lack of stringent pharmacophoric constraints allowed good binding modes even in noncognate proteins. Conversely, fragments showing high specificity were mostly characterized by the presence of polar functional groups at a conformationally constrained distance. This resulted in an enhanced ability to recognize native proteins within the data set. Overall, our docking assessment showed that reliable poses and scores for fragments can be achieved even using standard protocols. Nevertheless, there is room for improvement. Development of ad hoc written algorithms could increase the success rate and allow proper treatment of a wider variety of chemical interactions.

To the best of our knowledge, this is the first reported attempt to compile a data set of biomolecular complexes, extracted from the wwPDB, purposely designed to exclusively incorporate fragments that interact with biologically attractive proteins. The data set is publicly available in a ready-to-dock format at the following address: http://www.iit.it/en/drug-discovery-and-development/seraphic.html.

METHODS

Selection of Protein−Fragment Complexes. The data set of proteins complexed with fragment-like ligands was created as follows.

The PDBe database[53] (containing 65 968 entries on the 18th of June 2010) was queried with SQL (query 1, available as Supporting Information) using the Database Browser interface available at http://www.ebi.ac.uk/pdbe-as/pdbedatabase/PDBeDatabase.jsp. The SQL query limited the data set to entries that (i) were dated on or after January 2000, (ii) were associated with at least one PubMed publication, (iii) had a resolution of 2.5 A or better, (iv) contained proteins with at least 200 amino acids in the SEQRES record, and (v) contained ligands with at least 6 heavy atoms and a molecular weight equal or greater than 78 (the molecular weight of benzene) and less than 300. A subsequent query (query 2 in Supporting Information) was used to identify molecular entities in the PDBe that contained mutations as well as entities whose polymer type was DNA or RNA. A Perl script was used to remove those entries from the results of query 1 that appeared in the results of query 2. This eliminated proteins with mutations and polymers other than protein from our data set. The next step was to remove all protein−ligand pairs, where the ligand was considered to be commonly found in crystallization buffers (e.g., sulphates, glycol etc.). Three hundred sixty two HET codes

were removed and are listed in Supporting Information. This list includes the 361 unique HET codes listed by Hartshorn et al.[24] together with the ligand TLA (tartaric acid). We note that Hartshorn et al. list 420 codes, but some are duplicates. Next, we removed any proteins that were not classified in CATH,[23] contained multidomain chains, or contained multiple single-domain chains where the domains were different at the S95 level in CATH. Finally, we retained only those PDB entries for which an electron density map exists in the Uppsala Electron Density Server (EDS).[54] We did this by removing those entries that do not appear in the file http://eds.bmc.uu.se/eds/eds_holdings.txt. The resulting data set comprised 810 potential protein−ligand pairs. However, this included pairs where the protein chain and the ligand do not interact directly (such cases were removed in the final manual curation step described below).

The final part of the automated data set preparation was creation of a nonredundant protein subset. We used the CATH classification to assign proteins to homologous superfamilies (i.e., domains sharing the first four digits in the classification). We then kept one example from each superfamily either by selecting the highest resolution structure for a human protein if present or if no human proteins were included by selecting the highest resolution structure for the given superfamily. At this stage, however, we kept multiple occurrences of the same PDB entry if two ligands of interest were bound or if multiple chains could have interacted with the same ligand. The final step of the preparation was manual. It involved a visual inspection of the protein−fragment pairs and an evaluation of the fragment density fit. The final data set comprises 56 protein−ligand pairs, extracted from 53 PDB entries and representing 54 unique HET codes.

Preparation of Receptor Structures. Receptor coordinates were retrieved from the Protein Data Bank. Chains not involved in defining the binding region were deleted. After assigning the correct atom types according to a modified version of ECEPP/3 force field,[55] we added hydrogen atoms and missing heavy atoms. Zero occupancy side chains were optimized and assigned the lowest energy conformation. Tautomeric states of histidines and the positions of asparagine and glutamine side chain amidic groups were optimized to maximize the H-bond pattern. If available, histidines involved in direct interactions with ligands were assigned the tautomeric state reported in the primary literature. If not, the most energetically favorable state was calculated and assigned. Titratable residues were assigned standard protonation states at physiological pH. Polar hydrogen atoms were also optimized. Unless involved in specific contacts with both ligand and receptor, water molecules were deleted. After hydrogen optimization, cognate fragments were deleted from the complexes.

Preparation of Fragments. Ligand atomic coordinates were extracted from the crystallographic complexes. Bond order, tautomeric form, stereochemistry, and protonation state were assigned based on the primary literature description. Each ligand was assigned the MMFF force field[56] atom types and charges, and hydrogen atoms were added. Molecular properties for each fragment were calculated by means of QikProp.[57]

Binding Pockets. The boundaries of the binding pockets were defined taking into account the location of the fragment native bound pose. For each pocket, three descriptions of increasing size were provided, including all residues with at least one non-hydrogen atom within 3.5, 5, and 7.5 Å from a mesh representing the fragment molecular surface.[58] The three definitions yielded enclosing boxes that, on average, had sides of 18, 20, and 25 Å, respectively. In the pilot-VLS study we used the description at 5 Å from a mesh representing the fragment molecular surface. This yielded a total average volume of 8159 Å^3 per binding site. Pocket volumes were expressed as the volume of the largest envelope predicted within the pocket boundaries by the Pocketfinder, an algorithm that performs a Gaussian convolution of the Lennard−Jones potential.[59] The tolerance value was set equal to 4.6.

Docking of Fragments. The docking engine adopted was the Biased Probability Monte Carlo (BPMC) stochastic optimizer as implemented in ICM3.7.[45,56,60,61] The binding site was represented by precalculated 0.5 Å spacing potential grid maps, representing van der Waals potentials for hydrogens and heavy atoms, electrostatics, hydrophobicity, and hydrogen bonding. The van der Waals interactions were described by a smoothed form of the 6-12 Lennard−Jones potential, capping the repulsive contribution to 4 kcal/mol. A distance-dependent dielectric function was used (dielectric constant set equal to 1.0). To understand the consequences of the presence of bridging water molecules during the docking runs, we derived two sets of maps for each site considered (i.e., with and without solvent atoms). The initial thoroughness of the conformational search was calculated by an adaptive algorithm and scaled linearly, according to the number of rotatable bonds in the fragment. This standard value was systematically increased up to 10-fold, one unit at a time, to study the effects of extended sampling on docking accuracy. The binding energy was assessed with the standard ICM empirical scoring function.[62]

Fingerprints and Tanimoto Similarity. Molecular patterns were calculated and hashed into bitmaps according to the Daylight algorithm for fingerprint generation (Daylight Chemical Information Systems Inc., Laguna Niguel, CA) as implemented in ICM3.7 (Molsoft LLC, San Diego, CA). Similarity between fragments was reported in terms of the Tanimoto distance, T. T between molecules m_1 and m_2 is expressed by the formula $T = 1 - (c/(a + b + c))$, where c counts the bits on in m_1 and m_2, a counts the bits on in m_1 but not in m_2, and b counts the bits on in m_2 but not in m_1. T spans from 0 to 1, with 0 indicating that two molecules share the same fingerprint.

Electron Density Fit. Fragment placement in the crystallographic complexes was checked against the electron density data from the Uppsala EDS[54] using a density fit routine[63,64] implemented in the ICM scripting language.[65] In the region surrounding the fragment, electron density contributions were interpolated on a regularly spaced grid of 0.5 Å step. The fit was calculated from the average grid values in the centers of the atoms. For unambiguously solved molecules, the procedure returned values approaching the ideal fitting score of 1. Fragments that were assigned values ranging from −1 to 0.7 were considered to be incorrectly placed in the density.

Ligand-Binding Analysis. Each entry of the data set was analyzed using the recently developed BINANA software.[44] The PDB files were prepared in Maestro[66] using the Protein Preparation Wizard[67] routine to assign hydrogens, bond types, and likely protonation states. Then, fragments and receptors were saved as separate PDB-formatted files and processed through the python scripts prepare_ligand4.py and prepare_receptor4.py, respectively, as available in the AutoDockTools distribution.[68] Finally, the resulting files were given as input to BINANA, which returned human readable log files with a summary of the detected interactions (i.e., hydrophobic contacts, H bonds, salt bridges, and π interactions). Furthermore a VMD state file allowed for a

quick visual inspection of the results through VMD version 1.9 running on a 64-bit Linux workstation.[69] For the sake of clarity, in Table 2, interactions occurring between fragments and metal ions were detached from the salt bridges, as detected by BINANA, and reported separately.

■ ASSOCIATED CONTENT

Supporting Information. Two plots showing the distribution of all pairwise sequence identities; list of those frequent binders not considered to be biologically interesting fragments in this work; SQL queries used in the selection of protein−fragment complexes; heat map showing the LE analysis on the pilot-VLS; plot illustrating the distribution of selected fragments in the space of the first two principal components of standard physicochemical descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +39-010-71781576 (A.D.F.); +39-051-2099735 (A.C.). Fax: +39-010-7170187 (A.D.F.); +39-051-2099734 (A.C.). E-mail: angelo.favia@iit.it (A.D.F.); andrea.cavalli@unibo.it (A.C.).

■ ACKNOWLEDGMENT

The authors thank J. Durrant for sharing the data used to validate BINANA and the IIT computational platform initiative for providing computer time. The authors thank Grace Fox for proofreading the manuscript.

■ ABBREVIATIONS

SERAPhiC, selected fragment protein complexes; FB, fragment based; SPR, surface plasmon resonance; NMR, nuclear magnetic resonance; ITC, isothermal titration calorimetry; MM-GBSA, molecular mechanics-generalized Born surface area; QM, quantum mechanics; IFP, interaction fingerprint; SIFt, structural interaction fingerprint; wwPDB, Worldwide Protein Data Bank; MSCS, multiple solvent crystal structures; LE, ligand efficiency; PCA, principal component analysis; BPMC, biased probability Monte Carlo; EDS, Electron Density Server; VLS, virtual ligand screening

■ REFERENCES

(1) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531–4.

(2) Warr, W. A., Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **2009**.

(3) Retra, K.; Irth, H.; van Muijlwijk-Koezen, J. E. Surface Plasmon Resonance biosensor analysis as a useful tool in FBDD. *Drug Discovery Today: Technol.* **2010**, *7*, e181–e187.

(4) Dalvit, C. NMR methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug Discovery Today* **2009**, *14*, 1051–7.

(5) Chessari, G.; Woodhead, A. J. From fragment to clinical candidate--a historical perspective. *Drug Discovery Today* **2009**, *14*, 668–75.

(6) Turnbull, W. B.; Daranas, A. H. On the value of c: can low affinity systems be studied by isothermal titration calorimetry?. *J. Am. Chem. Soc.* **2003**, *125*, 14859–66.

(7) Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.* **2005**, *48*, 403–13.

(8) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–8.

(9) Chen, Y.; Shoichet, B. K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2009**, *5*, 358–64.

(10) Dymock, B. W.; Barril, X.; Brough, P. A.; Cansfield, J. E.; Massey, A.; McDonald, E.; Hubbard, R. E.; Surgenor, A.; Roughley, S. D.; Webb, P.; Workman, P.; Wright, L.; Drysdale, M. J. Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *J. Med. Chem.* **2005**, *48*, 4212–5.

(11) Teotico, D. G.; Babaoglu, K.; Rocklin, G. J.; Ferreira, R. S.; Giannetti, A. M.; Shoichet, B. K. Docking for fragment inhibitors of AmpC beta-lactamase. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7455–60.

(12) Warner, S. L.; Bashyam, S.; Vankayalapati, H.; Bearss, D. J.; Han, H.; Mahadevan, D.; Von Hoff, D. D.; Hurley, L. H. Identification of a lead small-molecule inhibitor of the Aurora kinases using a structure-assisted, fragment-based approach. *Mol. Cancer Ther.* **2006**, *5*, 1764–73.

(13) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

(14) Whittaker, M.; Law, R. J.; Ichihara, O.; Hesterkamp, T.; Hallett, D. Fragments: past, present and future. *Drug Discovery Today: Technol.* **2010**, *7*, e163–e171.

(15) Sandor, M.; Kiss, R.; Keseru, G. M. Virtual fragment docking by Glide: a validation study on 190 protein-fragment complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1165–72.

(16) Kawatkar, S.; Wang, H.; Czerminski, R.; Joseph-McCarthy, D. Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 527–539.

(17) Favia, A. D.; Nobeli, I. Using chemical structure to infer biological function. In *Computational Approaches in Cheminformatics and Bioinformatics*; Bender, A., Rajarshi, G., Eds.; Wiley: New York, in press.

(18) Friedman, R.; Caflisch, A. Discovery of plasmepsin inhibitors by fragment-based docking and consensus scoring. *ChemMedChem* **2009**, *4*, 1317–26.

(19) Gleeson, M. P.; Gleeson, D. QM/MM as a tool in fragment based drug discovery. A cross-docking, rescoring study of kinase inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 1437–48.

(20) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–48.

(21) Chen, Y.; Pohlhaus, D. T. In silico docking and scoring of fragments. *Drug Discovery Today: Technol.* **2010**, *7*, e149–e156.

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

(23) Orengo, C. A.; Pearl, F. M.; Thornton, J. M. The CATH domain structure database. *Methods Biochem. Anal.* **2003**, *44*, 249–71.

(24) 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*, 726–41.

(25) Berman, H. M.; Westbrook, J. D. The impact of structural genomics on the protein data bank. *Am. J. Pharmacogenomics* **2004**, *4*, 247–52.

(26) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605–2611.

(27) Babaoglu, K.; Shoichet, B. K. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2006**, *2*, 720–723.

(28) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput.-Aided Mol. Des.* **2008**, *22*, 311–25.

(29) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–90.
(30) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–80.
(31) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–9.
(32) Wimmerova, M.; Mitchell, E.; Sanchez, J. F.; Gautier, C.; Imberty, A. Crystal structure of fungal lectin: six-bladed beta-propeller fold and novel fucose recognition mode for Aleuria aurantia lectin. *J. Biol. Chem.* **2003**, *278*, 27059–67.
(33) Vogt, J.; Perozzo, R.; Pautsch, A.; Prota, A.; Schelling, P.; Pilger, B.; Folkers, G.; Scapozza, L.; Schulz, G. E. Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. *Proteins* **2000**, *41*, 545–53.
(34) Bernacki, K.; Kalyanaraman, C.; Jacobson, M. P. Virtual ligand screening against Escherichia coli dihydrofolate reductase: improving docking enrichment using physics-based methods. *J. Biomol. Screen.* **2005**, *10*, 675–81.
(35) Favia, A. D.; Nobeli, I.; Glaser, F.; Thornton, J. M. Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. *J. Mol. Biol.* **2008**, *375*, 855–74.
(36) Hermann, J. C.; Ghanem, E.; Li, Y.; Raushel, F. M.; Irwin, J. J.; Shoichet, B. K. Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc.* **2006**, *128*, 15882–91.
(37) Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. Structure-based activity prediction for an enzyme of unknown function. *Nature* **2007**, *448*, 775–9.
(38) Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry* **2005**, *44*, 2059–71.
(39) Macchiarulo, A.; Nobeli, I.; Thornton, J. M. Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol.* **2004**, *22*, 1039–45.
(40) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–41.
(41) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2010**, *38*, D355–60.
(42) Gasteiger, J. Of molecules and humans. *J. Med. Chem.* **2006**, *49*, 6429–34.
(43) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–801.
(44) Durrant, J. D.; McCammon, J. A. BINANA: A novel algorithm for ligand-binding characterization. *J. Mol. Graphics Modell.* **2011**, *29*, 888–93.
(45) Totrov, M.; Abagyan, R. Protein-Ligand docking as an energy optimization problem. In *Drug-receptor thermodynamics: introduction and applications*; Raffa, R. B., Ed.; Wiley: Chichester, New York, 2001; pp 603−624.
(46) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60*, 325–32.
(47) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54*, 5422–31.
(48) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–44.
(49) Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 356–64.
(50) Morphy, R.; Rankovic, Z. Fragments, network biology and designing multiple ligands. *Drug Discovery Today* **2007**, *12*, 156–60.
(51) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–64.
(52) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–1.
(53) Velankar, S.; Alhroub, Y.; Alili, A.; Best, C.; Boutselakis, H. C.; Caboche, S.; Conroy, M. J.; Dana, J. M.; van Ginkel, G.; Golovin, A.; Gore, S. P.; Gutmanas, A.; Haslam, P.; Hirshberg, M.; John, M.; Lagerstedt, I.; Mir, S.; Newman, L. E.; Oldfield, T. J.; Penkett, C. J.; Pineda-Castillo, J.; Rinaldi, L.; Sahni, G.; Sawka, G.; Sen, S.; Slowley, R.; Sousa da Silva, A. W.; Suarez-Uruena, A.; Swaminathan, G. J.; Symmons, M. F.; Vranken, W. F.; Wainwright, M.; Kleywegt, G. J. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **2011**, *39*, D402–10.
(54) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240–9.
(55) Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* **1992**, *96*, 6472–6484.
(56) Halgren, T. A. Merck molecular force field. I -V. *J. Comput. Chem.* **1996**, *17*, 490–641.
(57) *QikProp*, version 3.3; Schrödinger, LLC: New York, 2010.
(58) Totrov, M.; Abagyan, R. The contour-buildup algorithm to calculate the analytical molecular surface. *J. Struct. Biol.* **1996**, *116*, 138–43.
(59) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752–61.
(60) Abagyan, R.; Frishman, D.; Argos, P. Recognition of distantly related proteins through energy calculations. *Proteins* **1994**, *19*, 132–40.
(61) Abagyan, R.; Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235*, 983–1002.
(62) Totrov, M.; Abagyan, R. Derivation of sensitive discrimination potential for virtual screening, RECOMB '99. *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France; ACM Press: New York, 1999; pp 37−38.
(63) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **2009**, *52*, 397–406.
(64) Abagyan, R.; Kufareva, I. The flexible pocketome engine for structural chemogenomics. *Methods Mol. Biol.* **2009**, *575*, 249–79.
(65) Abagyan, R.; Orry, A.; Raush, E.; Totrov, M. *ICM Manual 3.7*; Molsoft LCC: La Jolla, CA, 2010.
(66) *Maestro*, version 9.1; Schrödinger, LLC: New York, 2010.
(67) *Schrödinger Suite 2010 Protein Preparation Wizard, Epik*, version 2.1; Schrödinger, LLC: New York, 2010. *Impact*, version 5.6; Schrödinger, LLC: New York, 2010. *Prime*, version 2.2; Schrödinger, LLC: New York, 2010.
(68) Sanner, M. F. Python: a programming language for software integration and development. *J. Mol. Graphics Modell.* **1999**, *17*, 57–61.
(69) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (33−8), 27–8.