Part One
Data Sources

Data Mining in Drug Discovery, First Edition. Edited by Rémy D. Hoffmann, Arnaud Gohier, and Pavel Pospisil.

1
Protein Structural Databases in Drug Discovery

Esther Kellenberger and Didier Rognan

1.1
The Protein Data Bank: The Unique Public Archive of Protein Structures

1.1.1
History and Background: A Wealthy Resource for Structure-Based Computer-Aided Drug Design

The Protein Data Bank (PDB) was founded in the early 1970s to provide a repository of three-dimensional (3D) structures of biological macromolecules. Since then, scientists from around the world submit coordinates and information to mirror sites in the Unites States, Europe, and Asia. In 2003, the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, USA), the Protein Data Bank in Europe (PDBe) – the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) before 2009, and the Protein Data Bank Japan (PDBj) at the Osaka University formally merged into a single standardized archive, named the worldwide PDB (wwPDB, http://www.wwpdb.org/) [1]. At its creation in 1971 at the Brookhaven National Laboratory, the PDB registered seven structures. With more than 75 000 entries in 2011, the number of structures being deposited each year in PDB has been constantly increasing (Figure 1.1).

The growth rate was especially boosted in the 2000s by structural genomics initiatives [2,3]. Research centers from around the globe made joint efforts to overexpress, crystallize, and solve the protein structures at a high throughput for a reduced cost. Particular attention was paid to the quality and the utility of the structures, thereby resulting in supplementation of the PDB with new folds (i.e., three-dimensional organization of secondary structures) and new functional families [4,5].

The TargetTrack archive (http://sbkb.org) registers the status of macromolecules currently under investigation by all contributing centers (Table 1.1) and illustrates the difficulty in getting high-resolution crystal structures, since only 5% targets undergo the multistep process from cloning to deposition in the PDB.

Data Mining in Drug Discovery, First Edition. Edited by Rémy D. Hoffmann, Arnaud Gohier, and Pavel Pospisil.

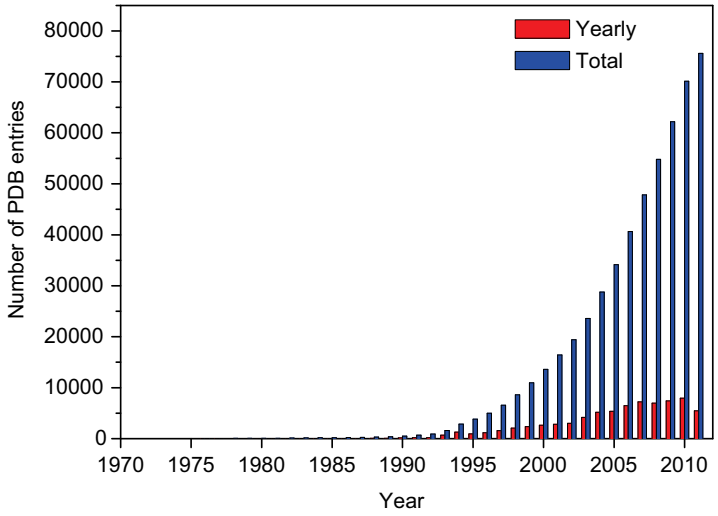


Figure 1.1 Yearly growth of deposited structures in the Protein Data Bank (accessed August 2011).

If only 450 complexes between an FDA-approved drug and a relevant target are available according to the DrugBank [6], the PDB provides structural information for a wealth of potential druggable proteins, with more than 40 000 different sequences that cover about 18 000 clusters of similar sequences (more than 30% identity).

Table 1.1 TargetTrack status statistics.

Status	Total number of targets	Relative to "cloned" targets (%)	Relative to "expressed" targets (%)	Relative to "purified" targets (%)	Relative to "crystallized" targets (%)
Cloned	192 735	100.0	—	—	—
Expressed	120 526	62.5	100.0	—	—
Soluble	35 436	18.4	29.4	—	—
Purified	45 105	23.4	37.4	100.0	—
Crystallized	14 472	7.5	12.0	32.1	100.0
Diffraction-quality crystals	7059	3.7	5.9	15.7	48.8
Diffraction	7522	3.9	6.2	16.7	52.0
NMR assigned	2262	1.2	1.9	5.0	—
HSQC	3409	1.8	2.8	7.6	—
Crystal structure	4953	2.6	4.1	11.0	34.2
NMR structure	2136	1.1	1.8	4.7	—
In PDB	8618	4.5	7.2	19.1	45

Accessed August 2011.

1.1.2 Content, Format, and Quality of Data: Pitfalls and Challenges When Using PDB Files

1.1.2.1 The Content

The PDB stores 3D structures of biological macromolecules, mainly proteins (about 92% of the database), nucleic acids, or complexes between proteins and nucleic acids. The PDB depositions are restricted to coordinates that are obtained using experimental data. More than 87% of PDB entries are determined by X-ray diffraction. About 12% of the structures have been computed from nuclear magnetic resonance (NMR) measurements. Few hundreds of structures were built from electron microscopy data. The purely theoretical models, such as *ab initio* or homology models, are no more accepted since 2006. For most entries, the PDB provides access to the original biophysical data, structure factors and restraints files for X-ray and NMR structures, respectively. During the past two decades, advances in experimental devices and computational methods have considerably improved the quality of acquired data and have allowed characterization of large and complex biological specimens [7,8]. As an example, the largest set of coordinates in the PDB describes a bacterial ribosomal termination complex (Figure 1.2) [9]. Its structure determined by electron microscopy includes 45 chains of proteins and nucleic acids for a total molecular weight exceeding 2 million Da.

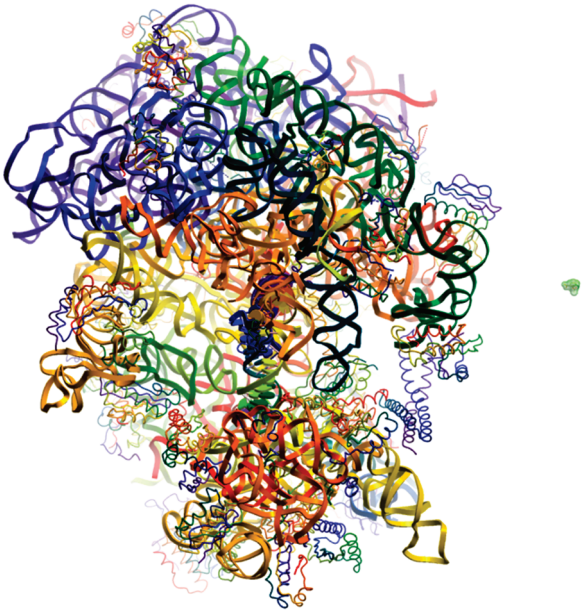


Figure 1.2 Comparative display of the largest macromolecule in the PDB (*Escherichia coli* ribosomal termination complex, PDB code 1ml5, left) and of a prototypical drug (aspirin, PDB code 2qqt, right).

To stress the quality issue, one can note the recent increase in the number of crystal structures solved at very high resolution: 90% of the 438 structures with a resolution better than 1 Å was deposited after year 2000. More generally, the enhancement in the structure accuracy translates into a more precise representation of the biopolymer details (e.g., alternative conformations of an amino acid side chain) and into the enlarged description of the molecular environment of the biopolymer, that is, of the nonbiopolymer molecules, also named ligands. Ligands can be any component of the crystallization solution (ions, buffers, detergents, crystallization agents, etc.), but it can also be biologically relevant molecules (cofactors and prosthetic groups, inhibitors, allosteric modulators, and drugs). Approximately 11 000 different free ligands are spread across 70% of the PDB files.

1.1.2.2 **The Format**

The conception of a standardized representation of structural data was a requisite of the database creation. The PDB format was thus born in the 1970s and was designed as a human-readable format. Initially based on the 80 columns of a punch card, it has not much evolved over time and still consists in a flat file divided into two sections organized into labeled fields (see the latest PDB file format definition at http://www.wwpdb.org/docs.html). The first section, or *header*, is dedicated to the technical description and the annotation (e.g., authors, citation, biopolymer name, and sequence). The second one contains the coordinates of biopolymer atoms (ATOM records), the coordinates of ligand atoms (HETATM records), and the bonds within atoms (CONECT records). The PDB format is roughly similar to the connection table of MOL and SD files [10], but with an incomplete description of the molecular structure. In practice, no information is provided in the CONECT records for atomic bonds within biopolymer residues. Bond orders in ligands (simple, double, triple, and aromatic) are not specified and the connectivity data may be missing or wrong. In the HETATM records, each atom is defined by an arbitrary name and an atomic element (as in the periodic table). Because the hydrogen atoms are usually not represented in crystal structures, there are often atomic valence ambiguities in the structure of ligands.

To overcome limits in data handling and storage capacity for very large biological molecules, two new formats were introduced in 1997 (the macromolecular crystallographic information file or mmCIF) and 2005 [the PDB markup language (PDBML), an XML format derivative] [11,12]. They better suit the description of ligands, but are however not widely used by the scientific community. There are actually few programs able to read mmCIF and PDBML formats, whereas almost all programs can display molecules from PDB input coordinates.

1.1.2.3 **The Quality and Uniformity of Data**

Errors and inconsistencies are still frequent in PDB data (see examples in Table 1.2). Some of them are due to evolution in time of collection, curation, and processing of the data [13]. Others are directly introduced by the depositors because of the limits in experimental methods or because of an incomplete knowledge of the chemistry and/or biology of the studied sample. In 2007, the wwPDB released a complete

Table 1.2 Common errors in PDB files and effect of the wwPDB remediation.

Description of errors	Impacted data	Status upon remediation
Invalid source organism	Annotation	Fixed
Invalid reference to protein sequence databases	Annotation	Fixed
Inconsistencies in protein sequences[a)]	Annotation	Fixed
Violation of nomenclature in protein[b)]	Structure	Fixed
Incomplete CONECT record for ligand residues	Structure	Partly solved
Wrong chemistry in ligand residues	Structure	Partly solved
Violation of nomenclature in ligand[c)]	Structure	Unfixed
Wrong coordinates[d)]	Structure	Unfixed

a) In HEADER and ATOM records.
b) For example, residue or atom names.
c) Discrepancy between the structure described in the PDB file and the definition in the Chemical Component Dictionary.
d) For example, wrong side chain rotamers in proteins.

remediated archive [14]. In practice, sequence database references and taxonomies were updated and primary citations were verified. Significant efforts have also been devoted to chemical description and nomenclature of the biopolymers and ligands. The PDB file format was upgraded (v3.0) to integrate uniformity and remediation data and a reference dictionary called the Chemical Component Dictionary has been established to provide an accurate description of all the molecular entities found in the database. To date, however, only a few modeling programs (e.g., MOE[1)] and SYBYL[2)]) make use of the dictionary to complement the ligand information encoded in PDB files.

The remediation by the wwPDB yielded in March 2009 to the version 3.2 of the PDB archive, with a focus on detailed chemistry of biopolymers and bound ligands. Remediation is still ongoing and the last remediated archive was released in July 2011. There are nevertheless still structural errors in the database. Some are easily detectable, for example, erroneous bond lengths and bond angles, steric clashes, or missing atoms. These errors are very frequent (e.g., the number of atomic clashes in the PDB was estimated to be 13 million in 2010), but in principle can be fixed by recomputing coordinates from structure factors or NMR restraints using a proper force field [15]. Other structural errors are not obvious. For example, a wrong protein topology is identified only if new coordinates supersede the obsolete structure or if the structure is retracted [16]. Hopefully, these errors are rare. More common and yet undisclosed structural ambiguities concern the ionization and the tautomerization of biopolymers and ligands (e.g., three different protonation states are possible for histidine residues).

1) Chemical Computing Group, Montreal, Quebec, Canada H3A 2R7.
2) Tripos, St. Louis, MO 63144-2319, USA.

To evaluate the accuracy of a PDB structure, querying the PDB-related databases PDBREPORT and PDB_REDO is a good start [15]. PDBREPORT (http://swift.cmbi.ru.nl/gv/pdbreport/) registers, for each PDB entry, all structural anomalies in biopolymers. PDB_REDO (http://www.cmbi.ru.nl/pdb_redo/) holds rerefined copies of the PDB structures solved by X-ray crystallography (Figure 1.3).

PDB_REDO

New PDB code

PDB entry 3rte

Structure

Spacegroup	I 4 2 2		
Cell dimensions	a: 122.441 Å	b: 122.441 Å	c: 155.050 Å
	α: 90.00°	β: 90.00°	γ: 90.00°
Resolution	2.10 Å		

Experimental data

Reflections	All: 32084	Test set: 1625 (5.1%)
Resolution range	38.72 Å	2.10 Å

R-values etc.

	From PDB header	Calculated from data	After conservative optimisation	After full optimisation
R	0.1620	0.1612	0.1636	0.1624
R-free	0.2070	0.2058	0.1975	0.1962
σR-free		0.0036	0.0035	0.0034
R-free Z-score		-0.89	2.31	2.32

WHAT_CHECK validation

	Original PDB entry	Conservatively optimised	Fully optimised
1st generation packing quality[1]	0.439	0.465	0.497
2nd generation packing quality[1]	-0.509	-0.443	-0.431
Ramachandran plot appearance[1]	-0.275	-0.036	-0.032
Chi-1/Chi-2 rotamer normality[1]	-0.900	0.110	0.555
Backbone conformation[1]	-0.345	-0.239	-0.271
Bond length RMS Z-score[2]	0.843	0.356	0.351
Bond angle RMS Z-score[2]	0.843	0.530	0.521
Total number of bumps[3]	32	26	23
Unsatisfied H-bond donors/acceptors[3]	16	18	20
Full WHAT_CHECK reports	Link	Link	Link

[1] Higher is better
[2] Should be lower than 1.000
[3] Fewer is better

Download

- Conservatively optimised structure (PDB | MTZ)
- Fully optimised structure (PDB | MTZ)
- YASARA scenes (for visualisation of the results)
- All files (compressed)
- PDB structure
- Structure factors

Figure 1.3 PDB_REDO characteristics of the 3rte PDB entry.

The quality issue was recently discussed in a drug design perspective with benchmarks for structure-based computer-aided methods [17–19]. A consensual conclusion is that the PDB is an invaluable resource of structural information provided that data quality is not overstated.

1.2 PDB-Related Databases for Exploring Ligand–Protein Recognition

The bioactive structure of ligands in complex with relevant target is of special interest for drug design. During the last decade, many databases of ligand/protein information have been derived from the PDB. Their creation was always motivated by the ever-growing amount of structural data. Each database however has its own focus, which can be a large-scale analysis of ligands and/or proteins in PDB complexes, or training and/or testing affinity prediction, or other structure-based drug design methods (e.g., docking). Accordingly, ligands are either thoroughly collected across all PDB complexes or only retained if satisfying predefined requirements. As a consequence, the number of entries in PDB-related databases ranges from a few thousands to over 50 000 entries. These databases also differ greatly in their content. This section does not intend to establish an exhaustive list. We have chosen to discuss only the recent or widely used databases and to group them according to their main purposes (Table 1.3).

1.2.1 Databases in Parallel to the PDB

The wwPDB contributors have developed free Web-based tools to match chemical structures in the PDB files to entities in the Chemical Component Dictionary; the Ligand Expo and PDBeChem resources are linked to the RCSB PDB and PDBe, respectively, and provide the chemical structure of all ligands of every PDB file [20,21]. A few other databases also hold one entry for each PDB entry. The Het-PDB database was designed in 2003 at the Nagahama Institute of Bio-Science and Technology to survey the nonbiopolymer molecules in the PDB and to draw statistics about their frequency and interaction mode [22]. It is still monthly updated and covers 12 000 ligands in the PDB. It revealed that the most repeated ligands in the PDB were metal ions, sugars, and nucleotides, all of which can be considered as part of the functional protein as a result of a posttraductional modification or as cofactors. Another important database was developed at Uppsala University to provide structural biologists with topology and parameters file for ligands [23]. This database named HIC-Up was maintained until 2008 by G. Kleywegt, who now leads the PDBe. Another useful service has been offered by the Structural Bioinformatics group in Berlin: the Web interface of the SuperLigands database allows the search for 2D and 3D similar ligands in the PDB [24]. The last update of SuperLigands was made in December 2009. Other PDB ligand warehouses have been developed during the last decade, but, like HIC-Up and SuperLigands, are not actively

Table 1.3 Representative examples of PDB-related databases useful for drug design.

Databases	Dates[a)]	Content	Web site
Repository of PDB ligands			
Ligand Expo	2004-	>13 000 different ligands Experimental and ideal coordinates of ligands (PDB, SD, mmCIF formats)	ligand-expo.rcsb.org
PDBeChem	2005-	>13 000 different ligands Experimental and ideal coordinates of ligands (PDB, SD, mmCIF formats)	www.ebi.ac.uk/pdbe/
HET-PDB	2004-	12 262 different ligands in 74 732 PDB files (August 2011) Navigator only, no download	hetpdbnavi.nagahama-i-bio.ac.jp
HiC-Up	1997–2008	7870 different ligands (March 2008) Experimental and ideal coordinates of ligands in PDB format. Dictionary files (X-PLOR/CNS, O, TNT)	xray.bmc.uu.se/hiccup
SuperLigands	2005–2009	10 085 different ligands in 401 300 complexes Experimental coordinates of ligands in PDB and MOL formats	bioinformatics.charite.de/superligands/
Experimental binding affinities			
PDBBind	2004-	Affinity data for 7986 PDB complexes	http://www.pdbbind.org.cn
Binding MOAD	2005-	Affinity data for 4782 PDB complexes	www.bindingmoad.org
BindingDB	2001-	721 721 affinity data for 60 179 proteins and 316 172 ligands, including PDB complexes	www.bindingdb.org/bind
ChEMBL	2008-	>5 million affinity data for 8603 proteins and >1 million ligands, including PDB complexes	www.ebi.ac.uk/chembl
Structural description of protein-ligand complexes			
Relibase	2003-	Experimental coordinates of the complex (in PDB and MOL2 format) or of the isolated ligand (in SD and MOL2 format)	relibase.ccdc.cam.ac.uk
sc-PDB	2006-	9891 protein–ligand complexes with refined hydrogen atom positions Separate coordinates for ligands (SD and MOL2 format), protein (PDB and MOL2 format), and active site (MOL2 format)	bioinfo-pharma.u-strasbg.fr/scPDB/
PSMDB		5266 nonredundant protein–ligand complexes Separate coordinates for ligands (SD format) and proteins (PDB format)	compbio.cs.toronto.edu/psmdb

a) The year of database creation is that of relative primary publication. It is followed by the year of the database last updated (- indicates that the database is still updated).

maintained, since the RCSB PDB and the PDBe directly integrate most of their data or services.

1.2.2 Collection of Binding Affinity Data

A few databases collect binding affinities such as experimentally determined inhibition (IC_{50}, K_i) or dissociation (K_d) constant for PDB complexes. The larger ones are Binding MOAD, PDBbind, and BindingDB [25–27]. Both Binding MOAD and PDBbind were developed at the University of Michigan, and have in common the separation of biologically relevant PDB ligands from invalid ones, such as salts and buffers. Their focuses are however different. For example, PDBbind disregards any complex without binding data, whereas Binding MOAD groups proteins into functional families and chooses the highest affinity complex as a representative. BindingDB considers only potential drug targets in the PDB, but collects data for many ligands that are not represented in the PDB.

In all cases, data gathering implies the manual review of the reference publications in PDB files and, more generally, expert parsing of scientific literature. BindingDB also contains data extracted from two other Web resources, PubChem BioAssay and ChEMBL. PubChem BioAssay database at the National Center for Biotechnology Information (NIH) contains biological screening results. ChEMBL is the chemogenomics data resource at the European Molecular Biology Laboratory. It contains binding data and other bioactivities extracted from scientific literature for more than a million bioactive small molecules, including many PDB ligands.

Affinity databases were recently made available from two of the wwPDB mirror sites. The RCSB PDB Web site now includes hyperlinks to the actively maintained ones, BindingDB and BindingMOAD. The PDBe Web site communicates with ChEMBL.

1.2.3 Focus on Protein–Ligand Binding Sites

As already described, RCSB PDB and PDBe resources currently provide chemical description and 3D coordinates for all ligands in the PDB. They also provide tools for inspection of protein–ligand binding (Ligand Explorer at RCSB PDB and PDBe-Motifs at PDBe). But as already discussed in this chapter, PDB data are prone to chemical ambiguities and not directly suitable to finely describe nonbonded intermolecular interactions. Several initiatives aimed at the structural characterization of protein–ligand interactions at the PDB scale. Among the oldest one is Relibase that automatically analyzes all PDB entries, identifies all complexes involving non-biopolymer groups, and supplies the structural data with additional information, such as atom and bond types [28]. Relibase allows various types of queries (text searching, 2D substructure searching, 3D protein–ligand interaction searching, and ligand similarity searching) and complex analyses, such as automatic superposition

of related binding sites to compare ligand binding modes. The Web version of Relibase is freely available to academic users, but does not include all possibilities for exploration of PDB complexes.

If Relibase holds as many entries as PDB holds ligand–protein complexes, other databases were built using only a subset of the PDB information. For example, the sc-PDB is a nonredundant assembly of 3D structures for "druggable" PDB complexes [29]. The druggability here does not imply the existence of a drug–protein complex, but that both the binding site and the bound ligand obey topological and physicochemical rules typical of pharmaceutical targets and drug candidates, respectively. Strict selection rules and extensive manual verifications ensure the selection in the PDB of binary complexes between a small biologically relevant ligand and a druggable protein binding site. The preparation, content, and applications of the sc-PDB are detailed in Section 1.3.

Along the same lines, the PSMDB database endeavors to set up a smaller and yet most diverse data set of PDB ligand–protein complexes [30]. Full PDB entries are parsed to select structures determined by X-ray diffraction with a resolution lower than 2 Å, with at least one protein chain longer than 50 amino acids, and a noncovalently bound small ligand. The PDB file of each selected complex was split into free protein structure and bound ligand(s). The added value of PSMDB does not consists in these output structure files that contain the original PDB coordinates, but in the handling of redundancy at both the protein and ligand levels.

With the growing interest of the pharmaceutical industry for fragment-based approach to drug design [31], several applications focusing on individual fragments derived from PDB ligands have recently emerged. Algorithms for molecule fragmentation were applied to a selection of PDB ligands defining a library of fragment binding sites [32] to map the amino acid preference of such fragments [33] or to extract possible bioisosteres [34].

1.3
The sc-PDB, a Collection of Pharmacologically Relevant Protein–Ligand Complexes

We decided in 2002 to set up a collection of protein–ligand binding sites called sc-PDB, originally designed for reverse docking applications [35]. While docking a set of ligands to a single protein was already a well-established computational technique for identifying potentially interesting novel ligands, the reverse paradigm (docking a single ligand to a set of protein active sites) was still a marginal approach. The main difficulty was indeed to automate the setup of protein–ligand binding sites with appropriate attributes, such as physicochemical (e.g., ionization and tautomerization states) and pharmacological properties of the ligand. It was not our intention to cover all ligand–protein complexes in the PDB, but rather to compile a large and yet not redundant set of experimental structures for known or potential therapeutic targets that had been cocrystallized with a known drug/inhibitor/activator or with a small endogenous ligand that could be replaced by a drug/inhibitor/activator (e.g., sildenafil in phosphodiesterase-5 is an adenosine mimic).

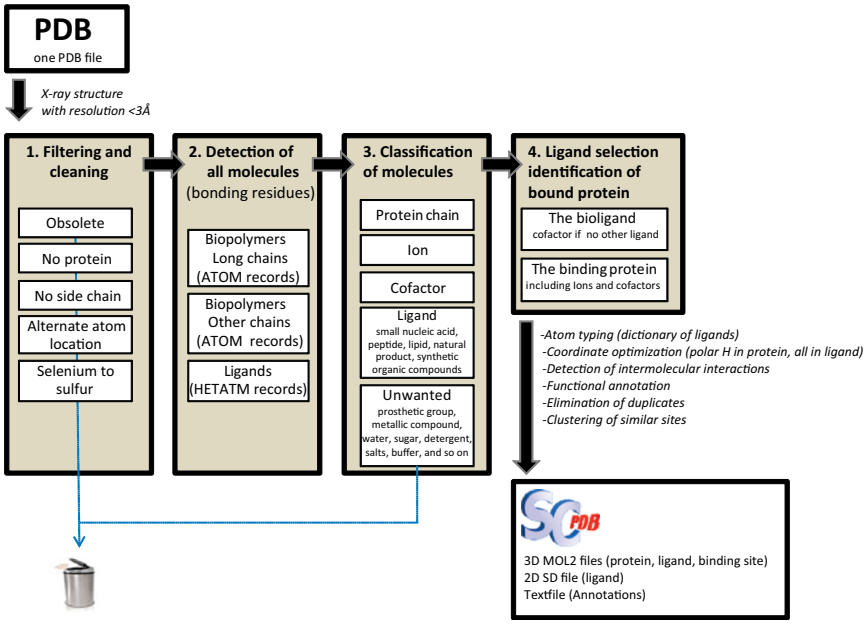


Figure 1.4 Flowchart to select sc-PDB entries from the PDB. Unwanted molecules at step 3 are identified using a dictionary or simple filters (based on ligand molecular weight, ligand surface area buried into the protein, number of amino acids close to the ligand, number of rings, and number of rotatable bonds of ligand). The bioligand in step 4 is the ligand that passes step 3 and maximizes the product of ligand molecular weight and surface area buried into the protein.

Selection rules as well as the applicability domain of the database have considerably evolved over time and are reviewed in the following sections.

1.3.1 Database Setup and Content

In brief, the selection scheme is made of simple and intelligible selection rules for the function and properties of the protein, the physicochemical properties of its ligand, and its binding mode (Figure 1.4).

The first publicly available version of the database was released in 2004 [35]. The database was named sc-PDB (acronym for *sc*reening the *P*rotein *D*ata *B*ank) (Table 1.4). At that time, it contained the atomic coordinates of proteins and their "druggable" binding sites. The protein was defined as all biopolymer chains, ions, and cofactors in the vicinity of the ligand. The binding site includes only the protein residues less than 6.5 Å away from the ligand. Noteworthy, all atoms were represented, including the hydrogen atoms not described in crystal structures. From 2005 onward, the sc-PDB has also provided the atomic coordinates of ligands. The ligand chemistry has been validated using an in-house dictionary, manually built from

Table 1.4 Annotation and available search options in the Web interface to the sc-PDB.

Object	Properties
PDB X-ray structure	PDB identifier Resolution Deposition date
Ligand	HET code Chemical structure Formula Molecular weight LogP LogS Polar surface area H-Bond donor count H-Bond acceptor count Number of rotatable bonds Number of rings Rule-of-five number of violations
Protein	Name EC number Uniprot accession number Uniprot name Source organism name Source organism taxonomy Source organism kingdom Mutant/wild type
Ligand binding site	Ion/cofactor Number of residues Number of nonstandard amino acids Number of chains Average B-factor Center of mass
Protein–ligand interactions	Number of hydrophobic interactions Aromatic face-to-face interactions Aromatic face-to-edge interactions H-Bond (donor in protein or ligand) Ionic interaction (cation in protein or ligand) Metal coordination Affinity data (K_i, K_d, IC_{50}, or pK_d) Ligand buried surface area

scratch then supplemented since 2007 by manually checked entries of the PDB Chemical Component Dictionary. The all-atoms representation of both partners of sc-PDB complexes have allowed us to refine the position of polar hydrogen atoms in the protein binding site and to compute an optimized pose of the bound ligand [29].

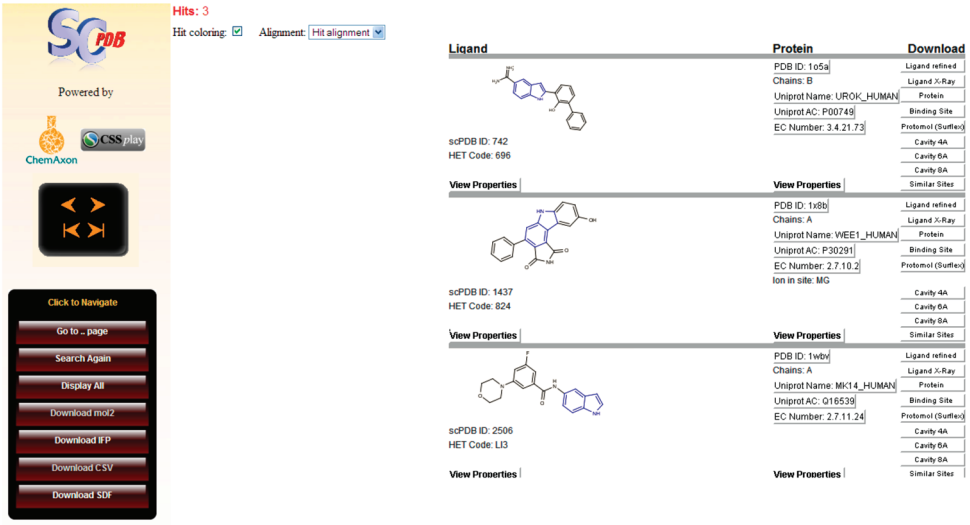


Figure 1.5 sc-PDB output for PDB protein–ligand complexes (3 hits) between an indole-containing ligand (blue substructure) of molecular weight <350 and a human kinase to which the ligand donates at least one hydrogen bond.

The sc-PDB is annually updated and regularly enriched with new information (ligand descriptors, binding mode encoded into an interaction fingerprint (IFP) [36], and cavity volume) and new functionalities (classification of similar binding sites [37]). A Web interface enables querying the database by combining requests about ligand chemical structures and properties, protein function and source organism, binding site properties, and ligand/protein binding properties (Figure 1.5).

The current version of the database contains 9891 entries corresponding to 3039 different proteins (according to protein sc-PDB name [37]) and 5505 different ligands (according to canonical SMILES strings). The sc-PDB protein space is redundant. There are 395 different proteins with more than 5 copies and single-copy proteins represent 55% of the database entries. Noteworthy is the complex nature of many proteins: a cofactor is bound to 219 proteins; calcium, magnesium, manganese, cobalt, zinc, or iron ions are found in 981 different proteins. No sc-PDB ligands are located at the interface of a protein–protein complex. The functional and species distribution of sc-PDB proteins reflects the bias in protein function space of the PDB itself, yet the sc-PDB is enriched in enzymes. The sc-PDB ligands space is also redundant and most prevalent ligands are cofactors and other nucleotides, which are also the most promiscuous ligands (e.g., more than 100 different protein targets for adenoside 5′-diphosphate or nicotinamide adenine dinucleotide). About 75% of the sc-PDB ligands is not primary bioorganic metabolites (nucleic acids, peptides, amino acids, sugars, or lipids) or their derivatives. Most of them pass the Lipinksi's rule of five (69%

with no violations and 20% with a single violation). The sc-PDB ligand space does not match that of commercial drugs because of a bias toward polar and flexible ligands. Finally, the sc-PDB ligand ensemble is not very diverse: for more than half of sc-PDB ligands, the ligand molecule is highly similar to at least one molecule in the pool of nonidentical ligands (with similarity evaluated by the Tanimoto coefficient, computed on feature-based circular 2D FCFP4 fingerprints, higher than 0.6).

1.3.2 Applications to Drug Design

1.3.2.1 Protein–Ligand Docking

The sc-PDB database has been developed for reverse docking applications [35] and is therefore an invaluable source for establishing large-scale docking benchmarks. Most validation studies, which flourished in the literature in the last decade, have been applied to a restricted set of a few hundred PDB targets [38–41] and in the best cases to a "clean" set of high-resolution protein structures in which erroneous PDB data (Table 1.2) have been removed [42]. In daily drug discovery programs, many targets under investigation do not obey such strict rules. Assessing the robustness of docking algorithms against a larger and more representative set of protein 3D structures is therefore of interest. The sc-PDB provides a unique source for such benchmarks since ligand, protein, and active site coordinates have been preprocessed and are ready for automated docking. When applied to a collection of 5681 complexes, Tietze and Apostoklasis reported with the GlamDock software [43] an accuracy (RMSD to the X-ray structure below 2.0 Å) significantly lower than that obtained with restricted protein sets with only 77% of sampling accuracy (RMSD of the best pose $<$2.0 Å) and 47% of scoring accuracy (RMSD of the top-ranked pose $<$2 Å). Along the same lines, we reported the accuracy of four docking algorithms in posing low molecular weight fragments into druggable sc-PDB binding sites and observed that ranking poses by a pure topological scoring function based on protein–ligand interaction fingerprints were much superior to poses by classical energy-based scoring functions [36].

Coming back to the seminal application for which the sc-PDB archive was initially developed (reverse docking), it appeared quite soon that the concept could be easily applied to a large and heterogeneous set of binding sites with a naïve target ranking scheme consisting of simple docking scores. Serial docking of four test ligands (biotin, methotrexate, 4-hydroxytamoxifen, and 6-hydroxy-1,6-dihydropurine ribonucleoside) to a collection of 2148 binding sites enabled recovering the known target(s) of the later ligands within the top 1% scoring entries, using the GOLD docking algorithm. These results were quite encouraging since these validated *per se* the reverse docking concept and notably the automated binding site setup protocol despite well-known insufficiencies regarding, for example, ionization/tautomerization of binding site residues as well as water-mediated ligand binding effects. These initial trials were applied to high-affinity ligands, which were relatively selective for very few targets. When applied to smaller and more permissive compounds

(e.g., AMP), a larger list of potential targets (top 5 to 10%) had to be selected to fish the correct protein targets [35]. The main reason was an inaccurate scoring of the "good" binding sites, which was not a real surprise with regard to the abundant literature about the limitations of fast scoring functions utilized in docking algorithms [19,44]. In order to overcome these severe limitations, alternative target ranking schemes independent of any energy calculation have been developed. One particular problem in docking-based target fishing is that the distribution of docking scores may be quite heterogeneous across different binding sites with diverse physicochemical properties. Therefore, score normalization according to either ligand and/or target properties is necessary to get rid of frequent target hitters [45–47]. Another promising approach consists in the conversion of protein–ligand coordinates (docking poses) into simple 1D IFPs [36]. Assuming that a virtual hit is more likely to be a true hit if it shares a similar target–ligand interaction profile with a known ligand, docking poses can be ranked by decreasing similarity of the IFP to that of the reference compound(s). Combining docking scores with IFP similarities allows removing many false positives (wrong targets with high docking scores), while still selecting the true targets in the final hit list [48].

1.3.2.2 **Binding Site Detection and Comparisons**

The sc-PDB provides, for each entry, all-atom Cartesian coordinates for the ligand, the target, and the binding site. By "binding site" we mean any monomer (amino acid, ion, cofactor, or prosthetic group) within 6.5 Å of any ligand heavy atom. Although the definition is conservative and excludes many potentially interesting pockets, it presents the advantage to favor cavities with well-described ligand occupancy. sc-PDB entries, therefore, can be used by cavity detection algorithms [49] to predict the most likely ligand binding sites and whether they are druggable or not, in other words, if the pocket could accommodate an orally available rule of five compliant drug-like molecule. When applied to 4915 sc-PDB protein structures, Volkamer *et al*. reported that the ligand is present in one of the three largest pockets in 90% of cases [50]. We used a grid-based cavity detection method (VolSite) to map cavity points with pharmacophoric properties of the closest protein atom, thus defining an ideal virtual ligand for each binding site (Figure 1.6).

Predicting the druggability of a given target from its three-dimensional structure is an intense field of research in order to reduce attrition rates in pharmaceutical discovery [51]. As druggability is by far more complex than the simple propensity of a particular protein cavity to accommodate high-affinity drug-like compounds, other terms, such as "bindability" [52] or "ligandability" [51] have been proposed recently, since they better capture target property ranges (cavity volume, polarity, and buriedness) known to be important for druggable targets [52–56]. Since these important properties are theoretically encoded in the aforementioned cavity site points, we investigated whether the present cavity descriptors might be suitable for predicting the ligandability of cavities from their 3D structures. A training set of 62 cavities (50% druggable and 50% undruggable) was assembled from literature [53,57] and the distribution of site point properties was given as input for a support vector machine (SVM) classifier. The best cross-validated classification model

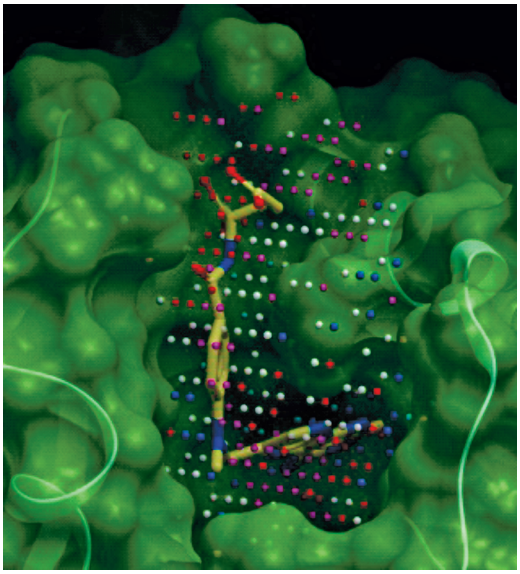
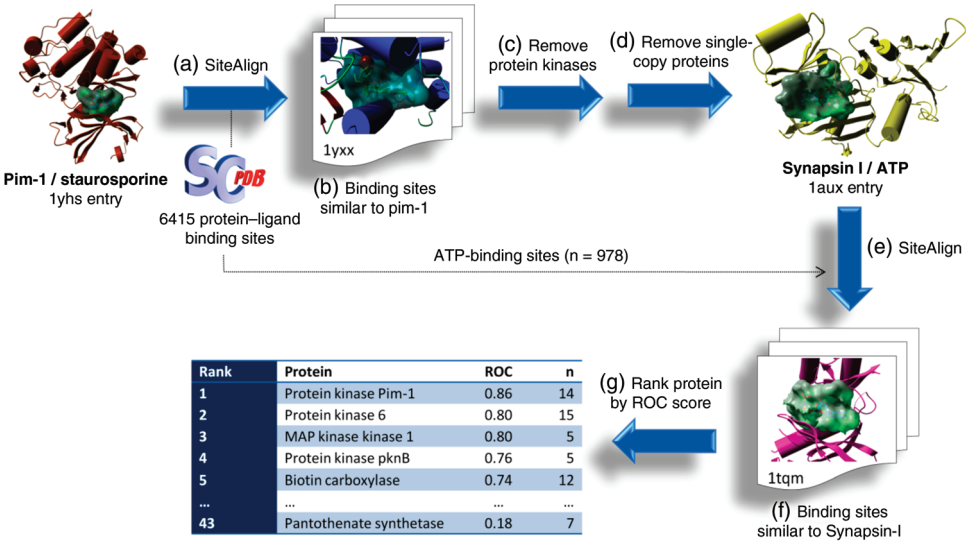


Figure 1.6 Detection and pharmacophoric annotation of VolSite cavity points in the X-ray structure of *Lactobacillus* dihydrofolate reductase (PDB code 4dfr). The cognate ligand (methotrexate, sticks) is shown in the binding site of the protein (green transparent surface). Cavity points are colored by pharmacophoric properties (H-bond acceptor and negative ionizable, red: H-bond donor and positive ionizable, blue: hydrophobe, white: aromatic, cyan: null, magenta).

achieves a very good accuracy of 80% and a Matthews correlation coefficient (MCC) of 0.62. Of course, larger sets of proteins of known (non)druggability are necessary to draw general conclusions, but the observed trend is quite promising and suggests that druggable target triage may be considered at an early level of drug discovery programs on condition that a high-resolution X-ray structure is available.

A second interesting application of the sc-PDB is the quantitative measure of its binding sites. Assuming that similar binding sites recognize similar ligands, comparing binding sites notably in the absence of 3D structure conservation permits identifying unexpected secondary targets for bioactive ligands. Several alignment-dependent or alignment-independent binding site comparison methods have been benchmarked on diverse collections of sc-PDB ligand binding sites [58–61] and have enabled the definition of global and local similarity thresholds for defining two sites as similar. Screening a library of binding sites for similarity to any given query is, therefore, possible and has already yielded the identification of an unexpected off-target (Synapsin I) for some but not all serine/threonine protein kinase inhibitors (Figure 1.7) [62].

Interestingly, only inhibitors of binding sites (cyclin-dependent kinase type 2, pim-1, and casein kinase II) predicted similar to that of Synapsin I were indeed found to bind to Synapsin I, sometimes with nanomolar affinities, whereas inhibitors of binding sites distant to that of Synapsin I (e.g., checkpoint kinase 1, protein kinase A, HSP-90α, DAG kinase, and DNA topoisomerase II) were not recognized by the enzyme [62].



Rank	Protein	ROC	n
1	Protein kinase Pim-1	0.86	14
2	Protein kinase 6	0.80	15
3	MAP kinase kinase 1	0.80	5
4	Protein kinase pknB	0.76	5
5	Biotin carboxylase	0.74	12
...	...	...	...
43	Pantothenate synthetase	0.18	7

Figure 1.7 Computational protocol used to detect local similarities between ATP-binding sites in pim-1 kinase and Synapsin I. The ATP-binding site in pim-1 kinase (occupied by the ligand staurosporine) is compared with SiteAlign [58] (step a) to 6415 binding sites stored in the sc-PDB database. Among the top scoring entries (step b), Synapsin I is the only protein not belonging to the protein kinase target family (step c) and present in numerous copies (step d). A systematic SiteAlign comparison (step e) of the ATP-binding site in Synapsin I with 978 other ATP-binding sites (from the sc-PDB) suggests that some but not all ATP-binding sites of protein kinases (steps f and g) resemble that of Synapsin I [62].

1.3.2.3 Prediction of Protein Hot Spots

The structural knowledge encoded by 3500 protein–ligand complexes in the sc-PDB has been used to derive a model able to discriminate, from simple 1D cavity fingerprints, 120 000 ligands interacting from 500 000 ligand-noninteracting protein atoms [63]. When applied to a novel complex, the model was able to predict with 70% accuracy the protein atoms that are likely to interact with a ligand and, therefore, prioritize protein structure-based pharmacophore queries specifically targeting these hot spots.

1.3.2.4 Relationships between Ligands and Their Targets

The sc-PDB data set offers the opportunity to delineate evolutionary relationships between ligands and their targets or binding sites. By examining the distribution patterns of sc-PDB ligands in the protein universe, Ji *et al.* reported that synthetic compounds (e.g., enzyme inhibitors) tend to bind to a single protein fold, whereas "superligands" (metabolites) are much more permissive and can be accommodated by more than 10 different protein folds [64]. Target fold promiscuity was almost found for ancestral ligands (e.g., nucleotide-containing metabolites) that appeared quite early in the evolution and behave as hubs of metabolic networks. Interestingly,

these ligands share common physicochemical properties (high flexibility and polarity) responsible for their promiscuity. Likewise, the analysis of cofactor usage (organic molecules and transition metal ions) by primitive redox proteins in the sc-PDB clearly shows that organic cofactors (NAD and NADP) are much more used than metals, probably because of the abundance of neutral residues at the border of the corresponding binding sites [65]. Finally, a survey of known interactions between phenolic ligands and their sc-PDB targets provides some explanations for the classically observed discrepancy between potent *in vitro* and moderate *in vivo* antioxidant properties of phenols [66]. A tight hydrogen bonding of phenolic moieties to many sc-PDB proteins suggests that reactive oxidative species (ROS) cannot be scavenged by phenols if they are already engaged in interactions with surrounding proteins.

Relationships between ligands and their targets could also be integrated in rational drug discovery programs. For example, retrieving from the sc-PDB, 171 diverse protein kinases cocrystallized with ATP competitors and aligning their binding sites led to the observation that crystal water patterns (position, hydrogen bond network to the kinase, and known inhibitor) were not necessarily conserved despite very high binding site similarities, thus suggesting novel avenues for optimizing the fine selectivity of kinases inhibitors [67]. By comparing the structure of unrelated targets binding to the same natural flavonoids, Quinn and coworkers introduced the concept of protein fold topology (PFT) [68] characterized by short stretches of not necessarily conserved secondary structures providing shared anchoring points to a common ligand. The concept was demonstrated for natural products binding to both biosynthetic enzymes and therapeutic targets and may explain why natural compounds are abundant among existing drugs [69].

1.3.2.5 **Chemogenomic Screening for Protein–Ligand Fingerprints**

In a recent report, Meslamani and Rognan describe a novel protein cavity kernel able to quantitatively measure the 3D similarity between two sc-PDB binding sites. A novel chemogenomic screening method based on a SVM was designed to browse the sc-PDB protein–ligand space and predict binary protein–ligand interactions from separate ligand and cavity fingerprints. The best SVM model was able to predict with a high recall (70%) and exquisite specificity (99%) and precision (99%) the binding of 14 117 external ligands to a set of 531 sc-PDB targets [70].

1.4 Conclusions

Exploiting structural knowledge on known protein–ligand complexes is a key step in the rational design of bioactive compounds. This knowledge has gained considerable value in the recent years, thanks to parallel endeavors of structural biologists and computational biologists/chemists to release an ever-increasing number of high-quality data. Many smart algorithms to parse and analyze the PDB have been described in the last couple of years with a large spectrum of applications ranging

from hit identification and optimization to massive ligand profiling against a large array of possible targets. With the expected better coverage of the therapeutic target space by the PDB in the coming years, we anticipate a significant boost of rational drug discovery and notably a better interplay between protein structure-based and ligand-centric methods.

References

1 Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, **35**, D301–D303.

2 Dessailly, B.H., Nair, R., Jaroszewski, L., Fajardo, J.E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B., and Orengo, C. (2009) PSI-2: structural genomics to cover protein domain family space. *Structure*, **17**, 869–881.

3 Nair, R., Liu, J., Soong, T.T., Acton, T.B., Everett, J.K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., Montelione, G.T., and Rost, B. (2009) Structural genomics is the largest contributor of novel structural leverage. *Journal of Structural and Functional Genomics*, **10**, 181–191.

4 Chandonia, J.M. and Brenner, S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

5 Brown, E.N. and Ramaswamy, S. (2007) Quality of protein crystal structures. *Acta Crystallographica Section D*, **63**, 941–950.

6 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., and Wishart, D.S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, **39**, D1035–D1041.

7 Joachimiak, A. (2009) High-throughput crystallography for structural genomics. *Current Opinion in Structural Biology*, **19**, 573–584.

8 Montelione, G.T. and Szyperski, T. (2010) Advances in protein NMR provided by the NIGMS Protein Structure Initiative: impact on drug discovery. *Current Opinion in Drug Discovery & Development*, **13**, 335–349.

9 Klaholz, B.P., Pape, T., Zavialov, A.V., Myasnikov, A.G., Orlova, E.V., Vestergaard, B., Ehrenberg, M., and van Heel, M. (2003) Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. *Nature*, **421**, 90–94.

10 Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., and Laufer, J. (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, **32**, 244–255.

11 Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D., and Fitzgerald, P.M.D. (1997) Macromolecular crystallographic information file. *Methods in Enzymology*, **277**, 571–590.

12 Westbrook, J., Ito, N., Nakamura, H., Henrick, K., and Berman, H.M. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.

13 Dutta, S., Burkhardt, K., Swaminathan, G.J., Kosada, T., Henrick, K., Nakamura, H., and Berman, H.M. (2008) Data deposition and annotation at the worldwide protein data bank. *Methods in Molecular Biology*, **426**, 81–101.

14 Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M., and Berman, H.M. (2008) Remediation of the protein data bank archive. *Nucleic Acids Research*, **36**D426–D433.

15 Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C., and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Research*, **39**, D411–D419.

16 Joosten, R.P. and Vriend, G. (2007) PDB improvement starts with data deposition. *Science*, **317**, 195–196.

17 Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T.M., Mortenson, P.N., and Murray, C.W. (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *Journal of Medicinal Chemistry*, **50**, 726–741.

18 Hawkins, P., Warren, G., Skillman, A., and Nicholls, A. (2008) How to do an evaluation: pitfalls and traps. *Journal of Computer-Aided Molecular Design*, **22**, 179–190.

19 Dunbar, J.B., Smith, R.D., Yang, C.-Y., Ung, P.M.-U., Lexa, K.W., Khazanov, N.A., Stuckey, J.A., Wang, S., and Carlson, H.A. (2011) CSAR benchmark exercise of 2010: selection of the protein–ligand complexes. *Journal of Chemical Information and Modeling*, **51**, 2036–2046.

20 Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M., and Westbrook, J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.

21 Golovin, A. and Henrick, K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.

22 Yamaguchi, A., Iida, K., Matsui, N., Tomoda, S., Yura, K., and Go, M. (2004) Het-PDB Navi.: a database for protein–small molecule interactions. *Journal of Biochemistry*, **135**, 79–84 [Erratum: *Journal of Biochemistry (Tokyo)*, 2004, 135 (5), 651.].

23 Kleywegt, G.J. and Jones, T.A. (1998) Databases in protein crystallography. *Acta Crystallographica Section D*, **54**, 1119–1131.

24 Michalsky, E., Dunkel, M., Goede, A., and Preissner, R. (2005) SuperLigands: a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics*, **6**, 122.

25 Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H.A. (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Research*, **36**, D674–D678.

26 Wang, R., Fang, X., Lu, Y., Yang, C.Y., and Wang, S. (2005) The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, **48**, 4111–4119.

27 Liu, T., Lin, Y., Wen, X., Jorissen, R.N., and Gilson, M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, **35**, D198–D201.

28 Hendlich, M., Bergner, A., Gunther, J., and Klebe, G. (2003) Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. *Journal of Molecular Biology*, **326**, 607–620.

29 Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *Journal of Chemical Information and Modeling*, **46**, 717–727.

30 Wallach, I. and Lilien, R. (2009) The protein–small-molecule database, a non-redundant structural resource for the analysis of protein–ligand binding. *Bioinformatics*, **25**, 615–620.

31 Rognan, D. (2012) Fragment-based approaches and computer-aided drug discovery. *Topics in Current Chemistry*, **317**, 201–222.

32 Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S.A., and Delfaud, F. (2009) Computational fragment-based approach at PDB scale by protein local similarity. *Journal of Chemical Information and Modeling*, **49**, 280–294.

33 Wang, L., Xie, Z., Wipf, P., and Xie, X.-Q. (2011) Residue preference mapping of ligand fragments in the Protein Data Bank. *Journal of Chemical Information and Modeling*, **51**, 807–815.

34 Wood, D.J., de Vlieg, J., Wagener, M., and Ritschel, T. (2012) Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *Journal of Chemical Information and Modeling*, **52**, 2031–2043.

35 Paul, N., Kellenberger, E., Bret, G., Muller, P., and Rognan, D. (2004) Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, **54**, 671–680.

36 Marcou, G. and Rognan, D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling*, **47**, 195–207.

37 Meslamani, J., Rognan, D., and Kellenberger, E. (2011) sc-PDB: a database for identifying variations and multiplicity of "druggable" binding sites in proteins. *Bioinformatics*, **27**, 1324–1326.

38 Verdonk, M.L., Berdini, V., Hartshorn, M.J., Mooij, W.T., Murray, C.W., Taylor, R. D., and Watson, P. (2004) Virtual screening using protein–ligand docking: avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, **44**, 793–806.

39 Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, **57**, 225–242.

40 Kontoyianni, M., McClellan, L.M., and Sokol, G.S. (2004) Evaluation of docking performance: comparative data on docking algorithms. *Journal of Medicinal Chemistry*, **47**, 558–565.

41 Perola, E., Walters, W.P., and Charifson, P.S. (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins*, **56**, 235–249.

42 Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T., Mortenson, P.N., and Murray, C.W. (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *Journal of Medicinal Chemistry*, **50**, 726–741.

43 Tietze, S. and Apostolakis, J. (2007) GlamDock: development and validation of a new docking tool on several thousand protein–ligand complexes. *Journal of Chemical Information and Modeling*, **47**, 1657–1672.

44 Ferrara, P., Gohlke, H., Price, D.J., Klebe, G., and Brooks, C.L., 3rd (2004) Assessing scoring functions for protein–ligand interactions. *Journal of Medicinal Chemistry*, **47**, 3032–3047.

45 Yang, L., Wang, K., Chen, J., Jegga, A.G., Luo, H., Shi, L., Wan, C., Guo, X., Qin, S., He, G., Feng, G., and He, L. (2011) Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome: clozapine-induced agranulocytosis as a case study. *PLoS Computational Biology*, **7**, e1002016.

46 Yang, L., Chen, J., and He, L. (2009) Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS Computational Biology*, **5**, e1000441.

47 Vigers, G.P. and Rizzi, J.P. (2004) Multiple active site corrections for docking and virtual screening. *Journal of Medicinal Chemistry*, **47**, 80–89.

48 Kellenberger, E., Foata, N., and Rognan, D. (2008) Ranking targets in structure-based virtual screening of 3-D protein libraries: methods and problems. *Journal of Chemical Information and Modeling*, **48**, 1014–1025.

49 Perot, S., Sperandio, O., Miteva, M.A., Camproux, A.C., and Villoutreix, B.O. (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today*, **15**, 656–667.

50 Volkamer, A., Griewel, A., Grombacher, T., and Rarey, M. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, **50**, 2041–2052.

51 Edfeldt, F.N., Folmer, R.H., and Breeze, A. L. (2011) Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, **16**, 284–287.

52 Sheridan, R.P., Maiorov, V.N., Holloway, M.K., Cornell, W.D., and Gao, Y.D. (2010) Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of Chemical Information and Modeling*, **50**, 2029–2040.

53 Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., and Huang, E.S. (2007)

Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, **25**, 71–75.
54 Hajduk, P.J., Huth, J.R., and Fesik, S.W. (2005) Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry*, **48**, 2518–2525.
55 Halgren, T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, **49**, 377–389.
56 Schmidtke, P. and Barril, X. (2010) Understanding and predicting druggability: a high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry*, **53**, 5858–5867.
57 Huang, N. and Jacobson, M.P. (2010) Binding-site assessment by virtual fragment screening. *PLoS One*, **5**, e10109.
58 Schalon, C., Surgand, J.S., Kellenberger, E., and Rognan, D. (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins*, **71**, 1755–1778.
59 Weill, N. and Rognan, D. (2010) Alignment-free ultra-high-throughput comparison of druggable protein–ligand binding sites. *Journal of Chemical Information and Modeling*, **50**, 123–135.
60 Totrov, M. (2011) Ligand binding site superposition and comparison based on Atomic Property Fields: identification of distant homologues, convergent evolution and PDB-wide clustering of binding sites. *BMC Bioinformatics*, **12** (Suppl. 1), S35.
61 Kasahara, K., Kinoshita, K., and Takagi, T. (2010) Ligand-binding site prediction of proteins based on known fragment–fragment interactions. *Bioinformatics*, **26**, 1493–1499.
62 Defranchi, E., Schalon, C., Messa, M., Onofri, F., Benfenati, F., and Rognan, D. (2010) Binding of protein kinase inhibitors to Synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One*, **5**, e12214.
63 Barillari, C., Marcou, G., and Rognan, D. (2008) Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *Journal of Chemical Information and Modeling*, **48**, 1396–1410.
64 Ji, H.F., Kong, D.X., Shen, L., Chen, L.L., Ma, B.G., and Zhang, H.Y. (2007) Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biology*, **8**, R176.
65 Ji, H.F., Chen, L., and Zhang, H.Y. (2008) Organic cofactors participated more frequently than transition metals in redox reactions of primitive proteins. *Bioessays*, **30**, 766–771.
66 Shen, L., Ji, H.F., and Zhang, H.Y. (2007) How to understand the dichotomy of antioxidants. *Biochemical and Biophysical Research Communications*, **362**, 543–545.
67 Barillari, C., Duncan, A.L., Westwood, I. M., Blagg, J., and van Montfort, R.L.M. (2011) Analysis of water patterns in protein kinase binding sites. *Proteins: Structure, Function, and Bioinformatics*, **79**, 2109–2121.
68 McArdle, B.M., Campitelli, M.R., and Quinn, R.J. (2006) A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. *Journal of Natural Products*, **69**, 14–17.
69 Kellenberger, E., Hofmann, A., and Quinn, R.J. (2011) Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. *Natural Product Reports*, **28**, 1483–1492.
70 Meslamani, J. and Rognan, D. (2011) Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *Journal of Chemical Information and Modeling*, **51**, 1593–1603.