Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing

Richard D. Smith, Jordan J. Clark, Aqeel Ahmed, Zachary J. Orban, James B. Dunbar Jr and Heather A. Carlson

Department of Medicinal Chemistry, *University of Michigan–Ann Arbor, 428 Church Street, Ann Arbor, MI 48109-1065, USA*

Correspondence to Heather A. Carlson: *carlsonh@umich.edu*

https://doi.org/10.1016/j.jmb.2019.05.024

Abstract

The goal of Binding MOAD is to provide users with a data set focused on high-quality x-ray crystal structures that have been solved with biologically relevant ligands bound. Where available, experimental binding affinities (K_a, K_d, K_i, IC_{50}) are provided from the primary literature of the crystal structure. The database has been updated regularly since 2005, and this most recent update has added nearly 7000 new structures (growth of 21%). MOAD currently contains 32,747 structures, composed of 9117 protein families and 16,044 unique ligands. The data are freely available on www.BindingMOAD.org. This paper outlines updates to the data in Binding MOAD as well as improvements made to both the website and its contents. The NGL viewer has been added to improve visualization of the ligands and protein structures. MarvinJS has been implemented, over the outdated MarvinView, to work with JChem for small molecule searching in the database. To add tools for predicting polypharmacology, we have added information about sequence, binding-site, and ligand similarity between entries in the database. A main premise behind polypharmacology is that similar binding sites will bind similar ligands. The large amount of protein–ligand information available in Binding MOAD allows us to compute pairwise ligand and binding-site similarities. Lists of similar ligands and similar binding sites have been added to allow users to identify potential polypharmacology pairs. To show the utility of the polypharmacology data, we detail a few examples from Binding MOAD of drug repurposing targets with their respective similarities.

Introduction

Structure-based drug design has benefited from the creation of several databases which combine structural information from the Protein Data Bank (PDB) [1,2] with biochemical affinity [3–14]. These databases all have varying requirements for inclusion and provide users with a wide range of information regarding the proteins, ligands and/or the protein–ligand complexes. Early protein data sets were small enough to exist only as a list of relevant PDB IDs inside of their corresponding publication. As the amount of data utilized in these types of studies has increased from mere tens of structures to the hundreds or even thousands of structures employed in more modern publications, the list sizes are too large to be included in their main body-text. This has resulted in data sets presented as separate downloadable entities or even hosted on the web as publicly accessible tools. Publicly available resources are of unquestionable utility to the scientific community, so long as they are maintained regularly and transparently described in their original publication as to be reproducible and appropriately utilized.

Binding MOAD [6] was originally published in 2005 as a database of carefully curated, high-quality, protein–ligand crystal structures of biologically interesting small molecules. This database includes binding data for many of the ligand–protein pairs, curated from their primary citation. The database is accessible *via* the web at www.BindingMOAD.org. Data are presented to users on a per-structure basis, but the proteins are also grouped by various sequence-based cutoffs to facilitate finding similar structures. Different versions of the data set are available for download. These include a version with only the structures for which there exist curated binding data, as well as a fully compressed and zipped copy of the collective biological unit files for all entries. The database has been updated on a near-annual basis.

PDBbind [14] is the only true competitor to Binding MOAD, providing a similar collection of protein data. The entrance criteria are similar and the provided subsets of data showcase where the databases differ. The Binding MOAD data set falls somewhere between PDBbind's general set and refined set, as PDBbind allows for non-x-ray structures and structures with poorer than 2.5-Å resolution in their general set [15]. The HiQ data set [16] available from Binding MOAD is not restricted to proteins with multiple complexes as in PDBbind's core set, and thus represents a larger number of protein targets. Both approaches of refining a stringent data set of high-quality structures are equally valid, users are encouraged to choose a data set based on the agreement between the curation criteria and the needs of their own experimental procedures. An update for Binding MOAD's HiQ set is anticipated for the latter half of 2019. The sc-PDB [7] is the most similar after PDBbind, but the pre-processed nature of its data set puts it into a docking/*in silico* pre-prep niche that sets itself apart. ChEMBL [17] and BindingDB [10] provide a tremendous amount of binding data for a significant number of protein targets. The majority of the ligand–target pairs in these two databases do not have corresponding experimentally determined structural data, resulting in a different category of database than Binding MOAD or PDBbind.

The rise in popularity, understanding, and availability of machine-learning techniques has resulted in an all-time high for production of new prediction-based algorithms, leading to even greater demand for data collections such as Binding MOAD [18,19]. Both Binding MOAD and the HiQ data set have been utilized by the community in training and benchmarking of various predictive algorithms and scoring functions. As an example, MOAD was recently utilized in training a method for assessing scoring function performance in binding affinity prediction [20].

Binding MOAD's large collection of small-molecule ligands and binding sites, combined with new features and presented data, allows for researchers to investigate more complex relationships, such as polypharmacology. Polypharmacology is when a small molecular ligand binds to multiple protein targets. Some of the practical applications of polypharmacology are drug repurposing and identifying the off-target binding behind drug side effects. Drug repurposing, or "repositioning," is the identification of new therapeutic uses for existing drugs [21]. Drug repurposing has emerged as an efficient and inexpensive approach, through which the early stages of drug development can be bypassed by discovering a new therapeutic area for an approved drug [21–23]. A computational technique commonly used in this repurposing is to identify similar ligands and binding sites with the hypotheses that (1) the chemical similarity between ligands of different targets can identify potential new targets for those molecules [24], and (2) the binding-site similarity of the protein targets can also be used to broaden the identification of new targets for those drugs. Web-based tools are beginning to emerge, which allow users to browse similar ligands and targets [21,22,25,26].

The most recent success in the area of drug repurposing has been in the development of *e*Repo-ORP to identify new drugs to combat rare orphan diseases [27]. The authors of eRepo-ORP generated models of the binding sites of drug–target pairs from DrugBank [28] using *e*Thread [29] and *e*FindSite [30] and compared them to models generated for the Orphanet [27] database of targets associated with orphan diseases using binding-site similarity determined by *e*MatchSite [31]. Their method identified 18,145 potential drug candidates for repurposing [27]. As an example of their success, a new inhibitor of KRAS was identified due to the binding-site similarity to PTK6, which bound the known drug vandetanib [27]. Naderi *et al*. [32] have used the same method as *e*Repo-ORP to generate *e*Model-BDB, which are binding site models generated from ligand–target pairs derived from BindingDB.

With polypharmacology and drug repurposing in mind, we have introduced ligand similarity data and binding-site similarity data to the Binding MOAD website. This work aims to update the community on details of the current structures in Binding MOAD along with additions and improvements made to the BindingMOAD.org website since the previous publication in 2015 [33]. We have migrated to Javascript applications of JChem, MarvinJS, and the NGL Viewer for performance and security reasons. We have also added data regarding sequence similarity, ligand similarity, and binding-site similarity. Lastly, we have expanded our collection to a total of 32,747 protein–ligand crystal structures, composed of 9117 protein families and 16,044 unique ligands.

Methods

Other protein–ligand databases such as ChEMBL and BindingDB cultivate their data in a "bottom-up" course, starting with the literature and available binding information for important ligands, and gathering structural data along the way if it is available. Since we are only interested in interactions where corresponding structural data exist, we operate along a "top-down" approach which starts with the PDB. We first import the entire PDB, remove inappropriate structures, and use the remaining structures to guide our literature searches in a systematic fashion. Since almost all protein structures are annotated with the authors' names and the appropriate reference, obtaining the reference for the literature portion of the search is straightforward.

Condensing the PDB and hand curation

Starting from the PDB (133,344 structures on 9/27/2017), our data pipeline assesses whether each protein structure is an appropriate entry for Binding MOAD (see Fig. 1). The specific contents and functions of this data pipeline have been detailed, previously [3,6,33]. The condensed description is as follows: Structures must be x-ray crystal structures of 2.5-Å resolution or better and contain at least one protein chain with a corresponding, non-covalently bound, biochemically relevant (valid) ligand. Structures emerging from the pipeline meeting these criteria are then hand curated for final entry into the database.

We emphasize that no protein–ligand structure is automatically processed from the PDB into our database without undergoing hand curation at *least* once. Literature citations for all final structures to be included in Binding MOAD are used to confirm the validity (biological relevance) of the ligands, as well as extract binding data. Our order of preference for affinity data is as follows: $K_d > K_i > IC_{50}$. Great care is taken to ensure that ligands entered into Binding MOAD are biochemically significant and are of relevant function in the crystal structure being considered (e.g., structures with only "invalid" crystallographic additives are not included in MOAD).

Addressing redundancy by sequence

Grouping proteins by similar sequence allows users to find multiple related structures, which makes various types of comparison and data set construction much easier. Enzyme classification (EC) numbers are used to group enzymes that perform similar catalytic reactions. Binding MOAD clustering was based on EC groupings in the past, but this method was abandoned for numerous reasons. The EC number listed in PDB files is not always correct, or present at all. In the latter case, filling in the missing data gaps is convoluted. However, most importantly, there still exists massive variation within ECs, so grouping into homologous protein families by sequence has proven to be more beneficial, straightforward, and reproducible. Structure sequences are compared using BLAST [34], and proteins are grouped into families by 90% sequence identity. Each family contains a "leader" complex, typically the complex with the tightest binding ligand, that is, the lowest K_i, K_d, or IC_{50} value, with K_d preferred. In cases where a family has no entry with binding data, complexes of ligand–protein or ligand–cofactor–protein are chosen over protein–cofactor complexes. When multiple complexes are available without affinity data, leaders are chosen by the following criteria in order:

1. Best resolution (complexes with ligands preferred over cofactor-only complexes)
2. Wild-type over structures with site mutations
3. Most recent deposition date
4. Factors such as R or R_{free} values
5. If all the above criteria are identical, the entries are likely from the same paper, which will be used to help in the tie-breaker

Addressing redundancy using unified binding sites

To compare binding sites across all the proteins in MOAD, we needed to create a robust definition of each binding site in each protein. Studies utilizing structural data will often define binding sites specific to each ligand-bound structure using a distance threshold to establish which residues are in contact with the bound ligand. As some protein targets contain massive binding sites, a significant amount of data is therefore missed by only considering residues in immediate contact with a bound small-molecule ligand of a single structure. More elegant approaches are available for proteins, which there exist multiple protein–ligand structures, to incorporate more data. Here, we introduce unified binding sites.

Unified binding sites represent the union of all protein residues in contact with *any* ligand of a given protein family (defined by protein sequence). The contacts are derived from the biounit files using a heavy-atom-to-heavy-atom threshold (4.5 Å in this case). In a family where there are N protein–ligand

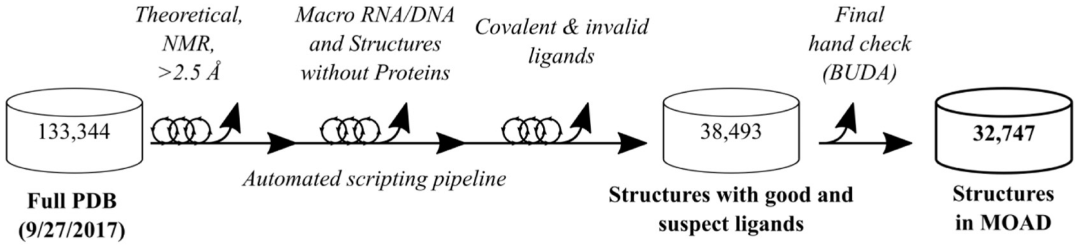


Fig. 1. Binding MOAD update process [6].

structures, the contacts from each of those *N* structures are assembled in one unified binding site, which describes the entire family. All ligands in the "leader" structure are identified, and independent union binding sites are constructed based on each ligand. This is a redundant set and ensures that each binding site is separate unless an overlap in residues exists. Sites are combined if there is an overlap of residues between the sites when creating the unified set of residues.

The most difficult aspect of assembling these unified binding sites resides in the protein numbering. Protein numbering is rarely always accomplished the same way in a group of more than a few structures. It is exceedingly difficult to identify and fix examples where numbering issues arise when using automated scripts for data processing. There are many examples of well-resolved, high-quality crystal structures that unfortunately suffer from multiple numbering disagreements. Therefore, we addressed this issue by renumbering protein structures prior to the assembly of unified binding sites. We stress that the renumbering was only used on the backend of the database to generate the unified binding-site data and compare the binding-site similarity. The original PDB numbering is used for any datafile a user might download.

To start, a similarity matrix of all protein chains in Binding MOAD was constructed using BLAST [34]. Using this similarity matrix, chains were then annotated for similarity (e.g., in dimer of heterodimers, chains A and C are often identical and chains B and D are often identical). These PDB ID/chain similarity indices are critical for the renumbering process, as knowing which chains within each PDB file are identical, and which chains are identical between files (different members of a same family with a 100% sequence identical chain). PDB SWS was used for renumbering templates [35]. PDB structures were then renumbered using the following framework:

1. If the PDB ID/chain combo is found in PDB SWS, renumber it accordingly.
2. If the PDB ID is found in PDB SWS but not for the current chain, use any sequence-identical chain within the same PDB ID that is found in PDB SWS.
3. If the PDB ID is not found in PDB SWS, use another structure with a 100% sequence-identical chain as the renumbering template.
4. If no structures in a homologous family are found in PDB SWS, check to see if their numbering already matches up.

In cases where multiple renumbering frameworks were provided by PDB SWS for a single homologous family, the mapping for the family leader was chosen and the whole family was renumbered in the same manner.

Ligand and binding-site similarities

The most popular and well-established measure of chemical similarity is the Tanimoto coefficient (Tc), which uses fingerprints for comparing two small-molecules. We have used PipelinePilot [36] and two different fingerprints, ECFP6 [37] and MDL [2] keys, to calculate the pairwise similarity between all ligands in Binding MOAD. Our experience with exploring chemical diversity has shown a cutoff of $\text{Tc} > 0.4$ for ECFP6 provides a reasonable definition of chemical similarity. However, our analyses determined that ECFP6 fingerprints yield many false negatives in the forms of $\text{Tc} < 0.4$ for similar molecules, or $\text{Tc} \neq 1$ for identical molecules. These results may be due to ECFP6 fingerprints' inherent sensitivity to SMILES strings (e.g., tautomers). Therefore, we have also used MDL keys in addition to ECFP6 and included all similarities with MDL score >0.8 (a customary cutoff also verified by manual inspection).

To further extend the ligand–target associations beyond ligand similarity, we applied binding-site similarity calculations to find all target pairs that share similar binding sites. These target–target associations are based on the idea that similar binding sites accommodate similar ligands. To compare binding sites, we first assembled unified binding sites to represent entire protein families as a single, condensed entity. The unified binding site is determined by all protein residues that contain a heavy atom within 4.5 Å of any ligand heavy atom within the family. Then, pairwise comparisons of unified binding sites between all combinations of family leaders in Binding MOAD were conducted. The presence of family leaders here is necessary as a physical manifestation of the unified binding site to be used in calculations, as a simple list of residues present in the binding sites is not sufficient.

Binding-site similarities between the unified binding sites of the family leaders were calculated using Alignment of Pockets (APoc). APoc [38] is an efficient program for large-scale structural comparison of protein pockets. We used the default parameters in APoc of 100 grid points for the pocket volume and at least 10 residues in a pocket. Only the unified binding sites were given to the program; therefore, global structure alignment was used. There were 14,916 unique unified binding sites in 9117 leaders using all available biounit files for each leader. Roughly, 111 million comparisons of pairs of binding sites were performed on 132 processors, simultaneously, which took 8 weeks of computational time. To extract the statistically significant binding-site associations, only *p*-values below 0.05 in Apoc were considered, which resulted in 3,510,682 target–target matches (32%). The *p*-value and PS_Score from the Apoc output are reported on BindingMOAD.org. It is already known that proteins with similar sequences have similar binding sites, so we only report binding-site matches for protein pairs that differ by more than 50% sequence similarity. As the unified binding sites are used in these calculations, the similarities between leaders inherently represent all structures contained within each of their families. Collapsible tables for both ligand similarity and binding-site similarity are located in the polypharmacology section for each complex (indexed by PDB ID) deposited in Binding MOAD.

Results and Discussion

The most recent update of Binding MOAD was derived from the version of the PDB extracted on September 27, 2017 (133,344 entries); a total of 32,747 valid protein–ligand complexes were obtained. Binding MOAD contains 16,044 unique, valid ligands within the 32,747 complexes. Comparatively, this updated data set contains 11,507 structures overlapping with PDBbind's [14] collection, representing 71.2% of their 16,151 total protein–ligand structures. These 11,507 overlapping structures contain 8538 unique ligands. In addition, 3385 of our 16,044 ligands are found as entries in Drugbank [28]. Figure 2 provides the distribution of the valid ligands in our collection by molecular weight. The ligands range from 4 to 278 heavy atoms, with an average molecular weight of 433 g/mol; an example of the average ligand is adenosine-5′-diphosphate (ADP), which has a molecular weight of 427 g/mol. Figure 2 shows that the number of large ligands (>500 g/mol) drops off quickly. The largest ligands are sugar chains, peptide chains (≤10 amino acids), and nucleic acid chains (≤4 nucleic acids).

Binding MOAD also contains 12,098 binding data across the 32,747 complexes (37% coverage of affinity data). These binding data are composed of 4128 K_d or K_a, 3788 K_i, and 4182 IC_{50} values. These binding affinities range over 16 orders of magnitude; Table 1 presents the range of binding values for each type of binding data, and the distribution of the three types of binding is further detailed in Fig. 3.

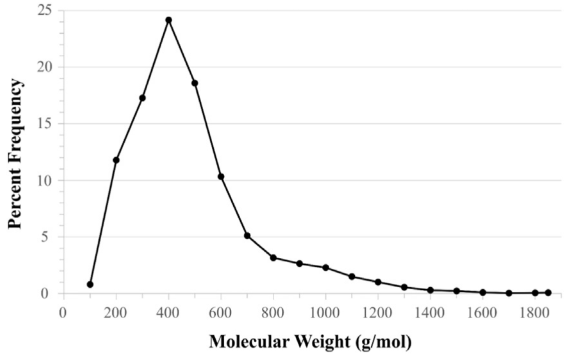


Fig. 2. Distribution of the current 16,044 unique ligands by molecular weight. The average ligand size in Binding MOAD is 433 g/mol. The largest are polysaccharides, peptides, and nucleic acid chains.

Table 1. The distribution of binding data within Binding MOAD

Binding data	Tightest	Lower quartile (nM)	Median	Upper quartile (μM)	Weakest (M)
K_d, K_a (as 1/	10 fM	110	2.51 μM	50.0	1.4

(continued on next page)

As mentioned previously, we are committed to the growth of Binding MOAD as a quality data resource in the community. Since being introduced in 2004, Binding MOAD has regularly expanded its collection with new data. Early updates brought in ~1500 new structures each year, but the rapid growth of the PDB has afforded us with many more structures in recent years. The growth of Binding MOAD is presented in Table 2 [33]. The delay in this most recent release stems from a 2-year lapse in funding.

Clustering Binding MOAD into homologous protein families

As noted above, the protein sequences of the entries in Binding MOAD are grouped into homologous protein families. Clustering at 90% sequence identity results in 9117 protein families, which is the default clustering, and individual data files for each of these families are available for download. In addition to families binned at 90% sequence identity, frequently it is necessary to think of a protein family at less strict cutoffs, like 70% or 50% sequence similarity. These families have been added to the data page for each of the entries in Binding MOAD. There are 7542 families when binned by 70% sequence similarity and 5768 families when binned by 50% sequence similarity. The clustering algorithm is a greedy algorithm, and not all entries within a family are necessarily within the similarity threshold to every other entry in the family.

New features and functionalities

Improved molecule and protein viewing

In order to improve the functionality of Binding MOAD and its web accessibility, programs that utilize JavaScript have been implemented for the viewing of binding sites and ligands. The *MarvinJS* [39] webservice from ChemAxon has replaced MarvinView to provide users with the ability to sketch small molecules and submit a query to JChem [40] (also from ChemAxon) for searching the small molecules in Binding MOAD. The current version utilizes

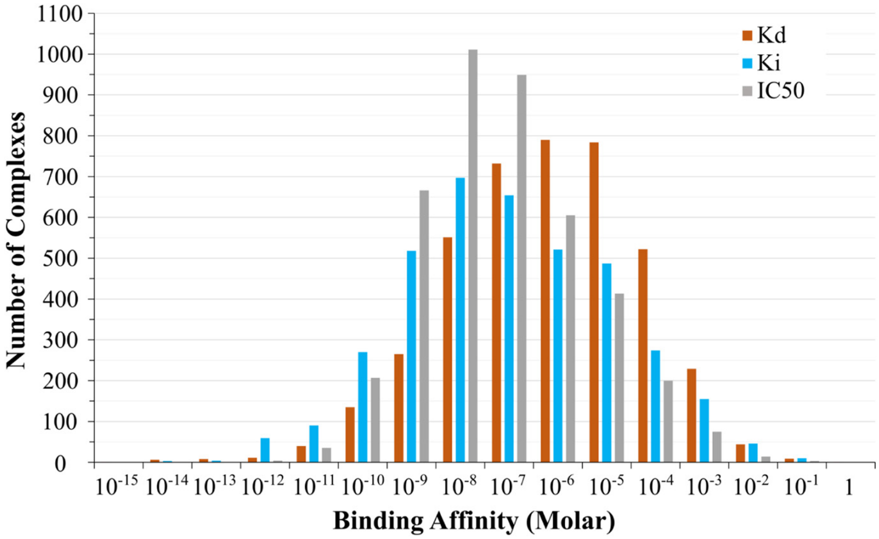


Fig. 3. The distribution of binding-affinity data within Binding MOAD. Data are available as K_d (orange), K_i (blue), or IC_{50} (gray). For this histogram, any K_a values were converted to K_d.

ChemAxon's MarvinJS webservice to stay up-to-date with newest releases of MarvinJS and avoid the need to update a license, since MarvinJS is free to academic institutions.

The *NGL Viewer* [41] is written in JavaScript and replaces the JMol viewer, which was HTML5-based. The NGL Viewer provides additional functionality to draw the molecular and solvent-accessible surfaces for individual protein and ligand residues. It also gives the user the ability to visualize the surface of interest. The solvent-accessible surface area of the ligand and unified binding sites (described below) are displayed in gray and blue, respectively, by default (Fig. 4).

Unified binding sites

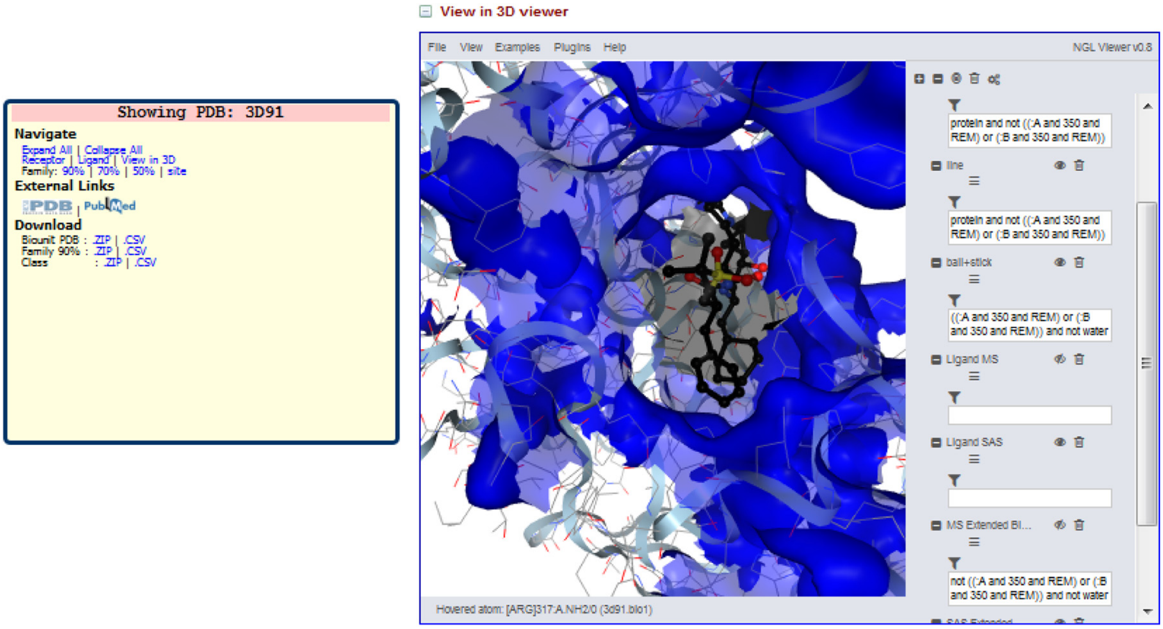
Many proteins bind a variety of ligands that differ in both size and chemical functionality, often resulting in different contacts with the protein. Using only one bound conformation to represent the binding site has the potential to miss important functional information, which leads to difficulty in exploring new chemical space. The unified binding site combines all binding-site residues from all structures in a protein family to obtain a more complete picture of the binding site. Visual representations of these binding sites have been added to the Binding MOAD website *via* the NGL viewer [41] (Fig. 4).

Ligand similarity

Ligand similarity is noted on each datapage in Binding MOAD. This information can be used to identify new targets for existing drugs in DrugBank [28]. Ligand similarities are presented in a table on the webpage associated with the structure it is bound. Figure 5 displays the ligand table for Nilotinib

Table 2. Growth Data for Binding MOAD (2004–2017)

Release (version, PDB download date)	PDB release size	Protein–ligand complexes	Protein families	Unique ligands	Binding affinity coverage
Initial release in 2004 [6]		5331	1780	2630	1375 (25.8%)
Prior to website in 2005		8250	2732	3932	2374 (28.8%)
1st (v2006, 12/31/2006) [3]	32,963	9836	3151	4665	2950 (30.0%)
2nd (v2007, 12/31/2007)	41,093	11,366	3583	5348	3452 (30.4%)
3rd (v2008, 12/31/2008)	48,168	13,138	4078	6210	4146 (31.6%)
4th (v2009, 12/31/2009)	56,466	14,720	4624	7064	4782 (32.5%)
5th (v2010, 12/31/2010)	65,344	16,948	5198	8140	5630 (33.2%)
6th (v2011, 12/31/2011)	74,594	18,764	5772	9048	6311 (33.6%)
7th (v2012, 12/31/2012)	84,566	21,109	6443	10,156	7284 (34.5%)
8th (v2013, 12/31/2013)	95,132	23,269	6960	11,173	8156 (35.0%)



Showing PDB: 3D91

Navigate

Expand All | Collapse All
Receptor | Ligand | View in 3D
Family: 90% | 70% | 50% | site

External Links

PDB | PubMed

Download

Biounit PDB : .ZIP | .CSV
Family 90% : .ZIP | .CSV
Class : .ZIP | .CSV

View in 3D viewer

Filter: protein and not ((:A and 350 and REM) or (:B and 350 and REM))

line

Filter: protein and not ((:A and 350 and REM) or (:B and 350 and REM))

ball+stick

Filter: ((:A and 350 and REM) or (:B and 350 and REM)) and not water

Ligand MS

Ligand SAS

MS Extended Bl...

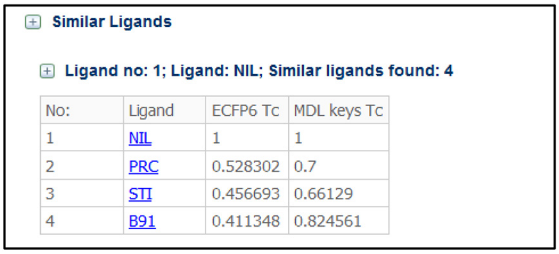
Filter: not ((:A and 350 and REM) or (:B and 350 and REM)) and not water

Hovered atom: [ARG]317:A.NH2/0 (3d91.blo1)

Fig. 4. Display of NGL Viewer in Binding MOAD with Aliskiren bound to Renin (PDB ID 3D91). The blue surface is the solvent accessible surface area of the unified binding site, and the gray transparent surface in the center is the solvent accessible surface area of the ligand shown in black sticks.

(HET code NIL) associated with PDB ID 3CS9's datapage in Binding MOAD. These tables can be collapsed or expanded as desired by the user.

Table 3 highlights three examples of polypharmacology that can be identified using ligand similarity. In each case, a known drug is similar to a ligand (HET group) in MOAD, which is crystallized to a target other than the traditional target. In each case, the secondary target has been confirmed in the literature, and in some instances, they have also been reported in DrugBank [28]. In the first example, simvastatin is a cholesterol-lowering drug that traditionally binds to HMG-CoA reductase. Simvastatin is crystallized bound to HMG-CoA reductase in PDB ID 1HW9, and it is chemically similar to the HET group AB6 (Tc = 0.51 by ECFP6), which is a lovastatin derivative bound to Integrin-α-L. The literature indicates that simvastatin is indeed an inhibitor of Integrin-α-L and is shown to induce apoptosis in lymphomas caused by the Epstein–Barr virus [28,43].



⊞ Similar Ligands

⊞ Ligand no: 1; Ligand: NIL; Similar ligands found: 4

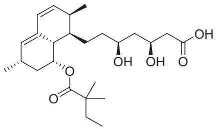
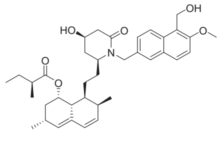
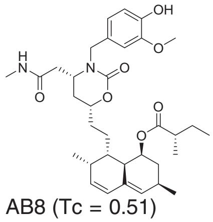
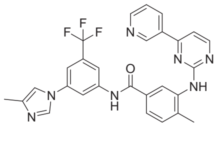
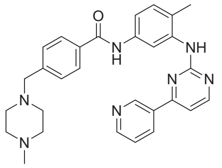
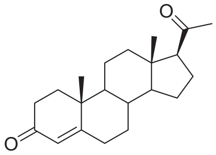
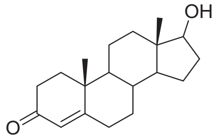
No:	Ligand	ECFP6 Tc	MDL keys Tc
1	NIL	1	1
2	PRC	0.528302	0.7
3	STI	0.456693	0.66129
4	B91	0.411348	0.824561

Fig. 5. Ligand similarity table for ligand NIL associated with PDB ID 3CS9 in Binding MOAD. The Tc for each match is given.

Binding-site similarity

We have examined over 40 published applications to compare two binding sites. The source codes for many of these applications were either unavailable or too computationally expensive for application on the scale of MOAD. A more complete list of binding-site comparison methods is found in a review by Jalencas and Mestres [44]. There were five methods that were able to successfully accommodate calculations of large data sets, Apoc [38], G-LoSA [45], ProBis [46], FuzCav [47], and PocketMatch [48]. Apoc was chosen due to its efficiency, as it is prohibitive to utilize all four programs on the entire data set simply to benchmark a best fit. However, we intend to add multiple binding-site comparison methods, the same way we use multiple ligand similarity measures. Pairwise calculations on 14,918 binding sites from 9117 families result in 111,183,872 total similarity calculations. Only protein families that have <50% sequence similarity and Apoc *p*-value < 0.05 are listed in the table to ensure unique protein targets are being identified. The data are displayed on the website as exemplified by the table for PDB ID 1OPK (Fig. 6). Complex 1OPK is ABL kinase binding an inhibitor P16, and the binding site matches are for other ATP-binding sites. All of these proteins bind ATP, ADP, and the inhibitor ANP. The matched binding sites are listed

Table 3. Examples of similar ligands binding to two very different targets, validated by DrugBank [28]

Drug (HET name) and Therapeutic Use	Traditional target and complex in MOAD	HET name of similar ligands in MOAD	Target of the similar ligand	Evidence that drug binds to alternate target
 Simvastatin (SIM): a cholesterol-lowering agent	HMG-CoA reductase PDB ID: 1HW9	 AAY (Tc = 0.48)  AB8 (Tc = 0.51) Both lovastatin derivatives	Integrin alpha-L (CD11 antigen-like family member A) PDB IDs: 1XDD and 1XDG	Integrin alpha-L-(simvastatin) in DrugBank [28].
 Nilotinib (NIL): treatment of imatinib-resistant chronic myelogenous leukemia	Tyrosine-protein kinase ABL1 PDB ID: 3CS9	 STI (Tc = 0.46) Imatinib	c-Kit Tyrosine kinase PDB ID: 1T46	Both targets in DrugBank [28]
 Progesterone (STR): steroid hormone	Progesterone receptor PDB ID: 1A28	 TES (Tc = 0.62) Testosterone	Androgen receptor PDB ID: 2AM9	Both targets in DrugBank [28] Binding between androgen receptor and progesterone is 5.71 nM [42]

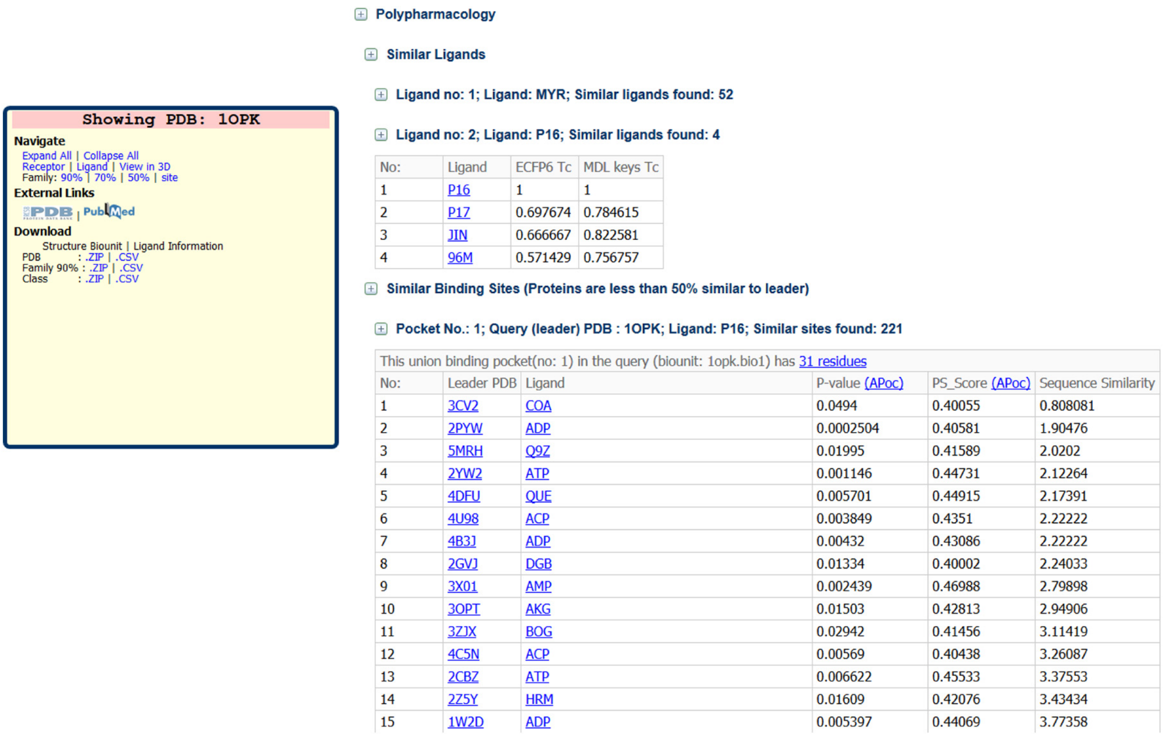
in a table sorted by sequence similarity, with the least similar sequences listed first so that the straightforward matches to closer homologs are listed last. This allows the user to identify the most unique matches first and the related homologs second.

To validate our approach, we examined the binding-site matches for known drug repurposing examples. Below we describe three examples of utilizing binding-site similarity in Binding MOAD to identify drugs that bind multiple targets.

Aliskiren is a Renin inhibitor that is commonly used in the treatment of hypertension [49]. Renin has been crystallized with Aliskiren (PDB ID 2V0Z) and shares less than 50% sequence similarity with the AIDS target HIV-1 protease (PDB ID 3T3C) despite both proteins being aspartic proteases. Using Apoc, we have identified that Renin and HIV-1 protease have similar binding sites (Apoc PS score: 0.52). It was recently confirmed that Aliskiren is a dual inhibitor of both Renin and HIV-1 protease [50]. Additional aspartic protease structures that show binding-site similarity larger than 0.5 include lysosomal aspartic protease (PDB ID 1LYB, PS score: 0.75), secreted aspartic protease (PDB IDs 2H6T, 1J71, and 2QZX, PS score: 0.70, 0.68, and 0.67 respectively), BACE-1 (PDB ID 4GID, 0.66), BACE-2 (PDB IDs 3ZKN, 3ZKI, and 3ZLQ 0.60, 0.59, and 0.59 respectively), HIV-2 protease (PDB IDs 5UPJ and 6UPJ, 0.53 and 0.52 respectively), EIAV protease (PDB ID 1FMB, 0.53) and FIV protease (PDB IDs 5FIV and 6FIV, 0.52), and HTLV-1 protease (PDB ID 3WSJ, 0.51). Note that these are all aspartic proteases. Tzoupis *et al*. [50] also performed docking studies with BACE-1 and HTLV-1 which suggest Aliskiren could inhibit these proteases as well.

Radicicol is an anti-tumor agent categorized as "experimental" in DrugBank. It is a known inhibitor of the chaperone Grp94 and was crystallized in complex with the protein (PDB ID 1U0Z) [51]. Topoisomerase VI, which shares only 16.5% sequence similarity, has a similar binding site to Grp94 (PS_Score = 0.53) and has been solved bound to radicicol (PDB ID 2HKJ) [52].

Imatinib was first marketed as a potent and specific Bcr-Abl kinase inhibitor in 2002 [53]. Although the initial structural characterization of its interactions was published, the actual crystal structure was not publicly deposited [54]. Since that time, numerous



Showing PDB: 1OPK

Navigate

Expand All | Collapse All
Receptor | Ligand | View in 3D
Family: 90% | 70% | 50% | site

External Links

PDB | PubMed

Download

Structure Biounit | Ligand Information
PDB : .ZIP | .CSV
Family 90% : .ZIP | .CSV
Class : .ZIP | .CSV

⊞ Polypharmacology

⊞ Similar Ligands

⊞ Ligand no: 1; Ligand: MYR; Similar ligands found: 52

⊞ Ligand no: 2; Ligand: P16; Similar ligands found: 4

No:	Ligand	ECFP6 Tc	MDL keys Tc
1	P16	1	1
2	P17	0.697674	0.784615
3	JIN	0.666667	0.822581
4	96M	0.571429	0.756757

⊞ Similar Binding Sites (Proteins are less than 50% similar to leader)

⊞ Pocket No.: 1; Query (leader) PDB : 1OPK; Ligand: P16; Similar sites found: 221

This union binding pocket(no: 1) in the query (biounit: 1opk.bio1) has 31 residues

No:	Leader PDB	Ligand	P-value (APoc)	PS_Score (APoc)	Sequence Similarity
1	3CV2	COA	0.0494	0.40055	0.808081
2	2PYW	ADP	0.0002504	0.40581	1.90476
3	5MRH	Q9Z	0.01995	0.41589	2.0202
4	2YW2	ATP	0.001146	0.44731	2.12264
5	4DFU	QUE	0.005701	0.44915	2.17391
6	4U98	ACP	0.003849	0.4351	2.22222
7	4B3J	ADP	0.00432	0.43086	2.22222
8	2GVJ	DGB	0.01334	0.40002	2.24033
9	3X01	AMP	0.002439	0.46988	2.79898
10	3OPT	AKG	0.01503	0.42813	2.94906
11	3ZJX	BOG	0.02942	0.41456	3.11419
12	4C5N	ACP	0.00569	0.40438	3.26087
13	2CBZ	ATP	0.006622	0.45533	3.37553
14	2Z5Y	HRM	0.01609	0.42076	3.43434
15	1W2D	ADP	0.005397	0.44069	3.77358

Fig. 6. Polypharmacology section of the datapage for complex PDB ID 1OPK on the Binding MOAD website.

x-ray crystal structures of Imatinib bound to Bcr-Abl have been deposited and four are present in Binding MOAD (PDB IDs 3MSS, 2HYY, 3K5V, 1IEP). Imatinib was later characterized by its interactions with p38α (PDB IDs 3HEC) [55]. Our calculations successfully identified similarity (PS score: 0.43) in the unified binding sites of Bcr-Abl and p38α, which may be expected as both proteins are kinases. It should be noted that imatinib has a case of extreme polypharmacology in that it also binds to human quinone reductase 2 [56]. Visual examination of the binding sites shows that they are very dissimilar (Apoc PS score = 0.30), but it is still identified through MOAD's ligand-similarity polypharmacology feature.

Here, we have presented just a few examples of previously observed polypharmacology for various targets in the literature. These examples were able to be identified *via* similarity of ligands, binding sites, or both. There surely exist many more examples which are still buried in the data, and every new update of extra structures will exponentially increase the number of possible target combinations to be investigated in the future.

Conclusions

We have detailed the further development and expansion of Binding MOAD. In the future, we aim to continue our annual updates to keep pace with the growth of the PDB. Binding MOAD has >32,000 hand-curated, protein–ligand x-ray crystal structures that contain ligands of biological relevance. Binding data are available for 37% of the entries, and this coverage has only increased with every update of the database. The value of Binding MOAD is not necessarily present in the quantity of its data, but more so in the quality. Maintaining this data quality is only achievable due to the considerable amount of effort placed in the update process and hand-curation. We have added similarity-based metrics to search the data set, both in terms of ligand similarity as well as protein similarity.

Our data sets are available online at http://www.BindingMOAD.org. This web-accessible resource is available to the research community, and our web interface also allows for users to contact us if they find any aspects of our curated data to be incorrect. Each structure's webpage includes details about ligands (both valid and invalid), available binding data, PDB ID for structural coordinates, EC class, homologous protein families with links to related structures at multiple sequence cutoffs (90%, 70%, 50%), a 3D visualization of the ligand bound in the unified binding site (using the NGL viewer [41]), as well as polypharmacology data presented as tables of ligand similarities and binding-site similarities.

Acknowledgments

This work has been supported in part by the National Institutes of Health (R01 GM124283). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Received 4 February 2019;
Received in revised form 13 May 2019;
Available online 22 May 2019

Keywords
protein-ligand complexes;
binding affinities;
ligand similarity;
binding-site similarity;
drug discovery

Abbreviations used:
PDB, Protein Data Bank; EC, enzyme classification; Tc, Tanimoto coefficient; APoc, Alignment of Pockets.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.

[2] , et al.. P.W. Rose, A. Prlic, A. Altunkaya, C. Bi, A.R. Bradley, C.H. Christie, et al., The RCSB protein data bank: integrative view of protein, gene and 3D structural information, Nucleic Acids Res. 45 (2017), D271-D81.

[3] M.L. Benson, R.D. Smith, N.A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, et al., Binding MOAD, a high-quality protein–ligand database, Nucleic Acids Res. 36 (2008), D674-D8.

[4] A. Bergner, J. Günther, M. Hendlich, G. Klebe, M. Verdonk, Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects, Biopolymers. 61 (2001) 99–110.

[5] A. Golovin, D. Dimitropoulos, T. Oldfield, A. Rachedi, K. Henrick, MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites, Proteins: Struct, Func. Bioinformatics. 58 (2005) 190–199.

[6] L. Hu, M.L. Benson, R.D. Smith, M.G. Lerner, H.A. Carlson, Binding MOAD (mother of all databases), Proteins: Struct. Func. Bioinformatics. 60 (2005) 333–340.

[7] E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata, D. Rognan, sc-PDB: an annotated database of druggable binding sites from the protein data Bank, J. Chem. Inf. Model. 46 (2006) 717–727.

[8] I. Kufareva, A.V. Ilatovskiy, R. Abagyan, Pocketome: an encyclopedia of small-molecule binding sites in 4D, Nucleic Acids Res. 40 (2012), D535–D40.

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

[10] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (2016) D1045–D1053.

[11] E. Michalsky, M. Dunkel, A. Goede, R. Preissner, Super-Ligands—a database of ligand structures derived from the Protein Data Bank, BMC BIOINFORMATICS. 6 (2005) 122.

[12] D. Puvanendrampillai, J.B. Mitchell, Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes, Bioinformatics. 19 (2003) 1856–1857.

[13] R. Wang, X. Fang, Y. Lu, S. Wang, The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures, J. Med. Chem. 47 (2004) 2977–2980.

[14] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, et al., PDB-wide collection of binding data: current status of the PDBbind database, Bioinformatics. 31 (2015) 405–412.

[15] Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li, et al., Forging the basis for developing protein–ligand interaction scoring functions, Acc. Chem. Res. 50 (2017) 302–309.

[16] J.B. Dunbar Jr., R.D. Smith, C.Y. Yang, P.M. Ung, K.W. Lexa, N.A. Khazanov, et al., CSAR benchmark exercise of 2010: selection of the protein–ligand complexes, J. Chem. Inf. Model. 51 (2011) 2036–2046.

[17] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, et al., ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012), D1100-D7.

[18] Q.U. Ain, A. Aleksandrova, F.D. Roessler, P.J. Ballester, Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening, Wiley Interdiscip Rev Comput Mol Sci. 5 (2015) 405–424.

[19] M. Wojcikowski, P.J. Ballester, P. Siedlecki, Performance of machine-learning scoring functions in structure-based virtual screening, Sci. Rep. 7 (2017), 46710.

[20] M.M. Xavier, G.S. Heck, M.B. Avila, N.M.B. Levin, V.O. Pintro, N.L. Carvalho, et al., SAnDReS: a computational tool for statistical analysis of docking results and development of scoring functions, Comb. Chem. High Throughput Screen. 19 (2016) 801–812.

[21] B. Karaman, W. Sippl, Computational drug repurposing: current trends, Curr. Med. Chem. 25 (2018) 1–19, https://doi.org/10.2174/0929867325666180530100332.

[22] M.R. Hurle, L. Yang, Q. Xie, D.K. Rajpal, P. Sanseau, P. Agarwal, Computational drug repositioning: from data to therapeutics, Clin. Pharmacol. Ther. 93 (2013) 335–341.

[23] Y.Y. Li, S.J. Jones, Drug repositioning for personalized medicine, Genome Med. 4 (2012) 27.

[24] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, et al., Predicting new molecular targets for known drugs, Nature. 462 (2009) 175–181.

[25] M. Awale, J.L. Reymond, The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data, J Cheminform. 9 (2017) 11.

[26] M. Awale, J.L. Reymond, Web-based tools for polypharmacology prediction, Methods Mol. Biol. 1888 (2019) 255–272.

[27] M. Brylinski, M. Naderi, R.G. Govindaraj, J. Lemoine, eRepo-ORP: exploring the opportunity space to combat orphan diseases with existing drugs, J. Mol. Biol. 430 (2018) 2266–2273.

[28] D.S. Wishart, Y.D. Feunang, A.C Guo, E.J. Lo, A. Marcu, J. R. Grant, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018), D1074-D82.

[29] M. Brylinski, D. Lingam, eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures, PLoS One 7 (2012), e50200.

[30] W.P. Feinstein, M. Brylinski, eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models, Mol Inform. 33 (2014) 135–150.

[31] M. Brylinski, eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models, PLoS Comput. Biol. 10 (2014), e1003829.

[32] M. Naderi, R.G. Govindaraj, M. Brylinski, eModel-BDB: a database of comparative structure models of drug–target interactions from the binding database, Gigascience. 7 (2018).

[33] A. Ahmed, R.D. Smith, J.J. Clark, J.B. Dunbar Jr., H.A. Carlson, Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures, Nucleic Acids Res. 43 (2015) D465–D469.

[34] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, et al., BLAST+: architecture and applications, BMC Bioinformatics. 10 (2009) 421.

[35] A.C. Martin, Mapping PDB chains to UniProtKB entries, Bioinformatics. 21 (2005) 4297–4301.

[36] Accelrys Software Inc, Pipeline Pilot 9.1.0.13 ed, Accelrys Software Inc., San Diego, CA, 2013.

[37] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.

[38] M. Gao, Skolnick J. APoc, Large scale identification of similar protein pockets, Bioinformatics. (2013) btt024.

[39] MarvinJS. 18.26.0 ed. Budapest, Hungary: ChemAxon Ltd.; 2018.

[40] S. Dakshanamurthy, N.T. Issa, S. Assefnia, A. Seshasayee, O.J. Peters, S. Madhavan, et al., Predicting new indications for approved drugs using a proteochemometric method, J. Med. Chem. 55 (2012) 6832–6848.

[41] A.S. Rose, P.W. Hildebrand, NGL Viewer: a web application for molecular visualization, Nucleic Acids Res. (2015) gkv402.

[42] B. van der Burg, R. Winter, H.Y. Man, C. Vangenechten, P. Berckmans, M. Weimer, et al., Optimization and prevalidation of the in vitro AR CALUX method to test androgenic and antiandrogenic activity of compounds, Reprod. Toxicol. 30 (2010) 18–24.

[43] H. Katano, L. Pesnicak, J.I. Cohen, Simvastatin induces apoptosis of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines and delays development of EBV lymphomas, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 4960–4965.

[44] X. Jalencas, J. Mestres, Identification of similar binding sites to detect distant polypharmacology, Mol Inform. 32 (2013) 976–990.

[45] H.S. Lee, W. Im, G-LoSA: an efficient computational tool for local structure-centric biological studies and drug design, Protein Sci. 25 (2016) 865–876.

[46] J. Konc, D. Janezic, ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins, Nucleic Acids Res. 40 (2012) W214–W221.

[47] N. Weill, D. Rognan, Alignment-free ultra-high-throughput comparison of druggable protein–ligand binding sites, J. Chem. Inf. Model. 50 (2010) 123–135.

[48] K. Yeturu, N. Chandra, PocketMatch: a new algorithm to compare binding sites in protein structures, BMC Bioinformatics. 9 (2008) 543.

[49] C.A. Sanoski, Aliskiren: an oral direct renin inhibitor for the treatment of hypertension, Pharmacotherapy. 29 (2009) 193–212.

[50] H. Tzoupis, G. Leonis, G. Megariotis, C.T. Supuran, T. Mavromoustakos, M.G. Papadopoulos, Dual inhibitors for aspartic proteases HIV-1 PR and renin: advancements in AIDS-hypertension-diabetes linkage via molecular dynamics, inhibition assays, and binding free energy calculations, J. Med. Chem. 55 (2012) 5784–5796.

[51] K.L. Soldano, A. Jivan, C.V. Nicchitta, D.T. Gewirth, Structure of the N-terminal domain of GRP94. Basis for ligand specificity and regulation, J. Biol. Chem. 278 (2003) 48330–48338.

[52] K.D. Corbett, J.M. Berger, Structural basis for topoisomerase VI inhibition by the anti-Hsp90 drug radicicol, Nucl Acids Res. 34 (2006) 4269–4277.

[53] R. Capdeville, E. Buchdunger, J. Zimmermann, A. Matter, Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug, Nat. Rev. Drug Discov. 1 (2002) 493–502.

[54] T. Schindler, W. Bornmann, P. Pellicena, W.T. Miller, B. Clarkson, J. Kuriyan, Structural mechanism for STI-571 inhibition of abelson tyrosine kinase, Science. 289 (2000) 1938–1942.

[55] H.V. Namboodiri, M. Bukhtiyarova, J. Ramcharan, M. Karpusas, Y. Lee, E.B. Springman, Analysis of imatinib and sorafenib binding to p38alpha compared with c-Abl and b-Raf provides structural insights for understanding the selectivity of inhibitors targeting the DFG-out form of protein kinases, Biochemistry. 49 (2010) 3611–3618.

[56] J.A. Winger, O. Hantschel, G. Superti-Furga, J. Kuriyan, The structure of the leukemia drug imatinib bound to human quinone reductase 2 (NQO2), BMC Struct. Biol. 9 (2009) 7.