J Comput Aided Mol Des (2013) 27:511–524
DOI 10.1007/s10822-013-9659-1

Protein pocket and ligand shape comparison and its application in virtual screening

Matthias Wirth · Andrea Volkamer · Vincent Zoete · Friedrich Rippmann · Olivier Michielin · Matthias Rarey · Wolfgang H. B. Sauer

Received: 15 March 2013 / Accepted: 12 June 2013 / Published online: 27 June 2013

Abstract Understanding molecular recognition is one major requirement for drug discovery and design. Physicochemical and shape complementarity between two binding partners is the driving force during complex formation. In this study, the impact of shape within this process is analyzed. Protein binding pockets and co-crystallized ligands are represented by normalized principal moments of inertia ratios (NPRs). The corresponding descriptor space is triangular, with its corners occupied by spherical, discoid, and elongated shapes. An analysis of a selected set of sc-PDB complexes suggests that pockets and bound ligands avoid spherical shapes, which are, however, prevalent in small unoccupied pockets. Furthermore, a direct shape comparison confirms previous studies that on average only one third of a pocket is filled by its bound ligand, supplemented by a 50 % subpocket coverage. In this study, we found that shape complementary is expressed by low pairwise shape distances in NPR space, short distances between the centers-of-mass, and small deviations in the angle between the first principal ellipsoid axes. Furthermore, it is assessed how different binding pocket parameters are related to bioactivity and binding efficiency of the co-crystallized ligand. In addition, the performance of different shape and size parameters of pockets and ligands is evaluated in a virtual screening scenario performed on four representative targets.

Keywords Protein binding sites · Molecular shape · Shape complementarity · Molecular recognition · Ligand efficiency · Virtual screening

Introduction

The identification of a small molecule that is able to modulate or block specific protein functions is one major goal in pharmaceutical research. Shape complementarity between ligand and its binding site is a condition for molecular recognition, but not alone sufficient. The binding of a small molecule to a receptor requires additional complementarity of electronic features [1]. However, even if electrostatic attraction can be significant over large distances, steric repulsion can counter any such interaction [2]. Therefore, one might rank the three properties size, shape, and electrostatics in this exact order: A molecule needs to have an appropriate size to enter the binding

Matthias Wirth and Andrea Volkamer contributed equally to this manuscript.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9659-1) contains supplementary material, which is available to authorized users.

M. Wirth · W. H. B. Sauer
Computational Chemistry, Merck Serono S.A. Geneva, Chemin des Mines 9, 1202 Geneva, Switzerland
e-mail: wsauer@bluewin.ch

M. Wirth (✉) · V. Zoete · O. Michielin
Swiss Institute of Bioinformatics, Molecular Modelling Group, UNIL Sorge-Bâtiment Génopode, 1015 Lausanne, Switzerland
e-mail: matthias.wirth@unil.ch

A. Volkamer · M. Rarey
Center for Bioinformatics, University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany
e-mail: volkamer@zbh.uni-hamburg.de

M. Rarey
e-mail: rarey@zbh.uni-hamburg.de

F. Rippmann
Global Computational Chemistry, Merck KGaA, Merck Serono, Frankfurter Str. 250, 64293 Darmstadt, Germany

cavity, it needs to be able to exhibit a comparable shape preventing clashes with the protein and to ensure optimal positioning of its functional groups in the molecular context to finally establish the necessary electrostatic complementarity. Many computational drug discovery approaches including molecular modeling and docking are based on a correct representation of these mechanisms and benefit from a better understanding of the recognition driving forces. Several computational studies have already been pursued to identify the importance of shape and chemical complementary between two binding partners [3–6]. Morris et al. [7], e.g., claimed that a large proportion of recognition potency resides in a tight shape fit between the two reaction partners. Many approaches exist that describe the molecular shape of small molecules in different granularities [8–10]. These abstractions are often used to virtually screen through large compound collections, e.g., by calculating the similarity between a reference compound and other chemical entities. A comprehensive perspective on the use of molecular shape approaches in medicinal chemistry has been published by Nicholls et al. [6].

Shape comparison methods, such as ROCS [11] and SQW [12], calculate a similarity value based on 3D coordinates between a query ligand and a database [9]. ROCS, e.g., computes a volume overlap of the molecules being compared and has been used successfully in many screening experiments. Another such method is the Ultra Fast Shape comparison (USR) [13], where molecules are compared based on the moments of four distributions of atomic distances. Due to its simple calculation, this method outperforms other shape descriptions with respect to computing time requirements. Another prominent approach is using Normalized Principal Moments of Inertia Ratios (NPRs) [14]. Besides not requiring a superposition of the input molecules, NPRs are independent of size, show low computational complexity, and allow the projection onto a finite triangular space that can easily be visualized. Based on the position of a data point in this NPR space, it can be assigned as being rod-, disc-, or sphere-like. In the context of multiple-scaffold versus single-scaffold combinatorial compound libraries, the idea of a pocket-shape focused space was introduced [14]. Akritopoulu-Zanze et al. [15] used NPRs to compare distributions of rule-of-five compliant compounds from the MDDR database and the corporate compound collection of Abbott. A drug-like shape space was derived and used to actively bias compound selection for HTS screening as well as for compound acquisition. In a recent study, the shape distribution in NPR space of small molecules originating from various large data sets was analyzed [16]. Globular ligand shapes were very rarely observed over all data sets.

Equally to the analysis of small molecular shapes, several studies explore the shapes of protein binding pockets. Already in 1998, a study of the anatomy of protein pockets revealed that the size and shape space of ligand binding sites is manifold [17]. Found shapes spread from simple spheres to more complicated shapes like curved grooves of several interconnected subpockets. Sonavane and Chakrabarti [18] describe the shape of a pocket by an estimated surface to volume ratio of a cavity relative to that of a sphere having the same volume as the cavity. They report a large number of globular pockets, but almost exclusively of small size. Weisel et al. [19] provide an overview about pocket architectures in a selected set of 623 co-crystallized complexes. In their work, they identify recurrent pocket topology patterns with the majority being elongated and containing one or more subbranches.

In addition to performing ligand-based shape comparisons between reference ligands and potential active molecules, if available, drug discovery projects can heavily benefit from structural information of the targeted protein. Various docking tools are available to algorithmically fit small molecules into the protein context and to assess their complementarity by scoring potential interactions. As this approach is computationally expensive, especially in the lead identification phase, a reduction of the search space through wise filtering can be of great benefit. Here, a detailed comparison of ligand and pocket sizes and shapes could be helpful as a first filtering step with the aim to reduce the size of the initial compound set to dock. In this context, some recent effort towards a direct comparison of ligand and pocket shape was promoted. Spherical harmonics have been investigated as shape representation in a work by Kahraman et al. [4], revealing that binding pockets are more flexible in their shapes than their respective ligands. Furthermore, the study showed that binding pocket and ligand volumes differ on average by a factor of three. Pérot et al. [20] observed in a case study on 56 complexes of the Astex set a correlation between pocket and ligand volumes and highlighted several cases of tight shape complementarity.

In this work, we present an analysis of ligand and pocket shapes using selected data from the sc-PDB [21]. Pockets and subpockets are predicted with DoGSite [22], followed by the calculation of shape properties [23]. Ligand and pocket shapes are approximated by ellipsoidal and moment of inertia main axes. The present article consists of three evaluation parts: First, an extensive analysis on pairwise combinations of binding pocket and corresponding ligand with respect to their shapes is performed. In accordance to the lack of globular ligand shapes in nature [16], globular ligand binding sites are found to be underrepresented. Shapes of occupied binding pockets are generally observed to be more globular than their containing ligand. When considering all cavities detected by DoGSite—including those without bound ligand and disabling a minimum volume threshold for predicted pockets—globular shapes are detected predominantly for small pockets. Furthermore, the respective ligand and pocket shape overlap is analyzed; pockets exhibit an average coverage of one third, while subpockets are on average covered by one half. The second part of this article addresses the relation between shape complementarity and bioactivity. In addition, it is analyzed whether information about the binding pocket of a particular target can be used to give an insight into the maximal achievable efficiency of a novel compound binding to this pocket. Finally, we analyze the performance of shape and size parameters retrieved from binding pocket and co-crystallized ligand in a virtual screening context. It is assessed if these parameters can be used to filter out compounds prior to screening, while preserving highly ranked actives.

Data preparation

Shape analysis data set

Version 2011 of the sc-PDB [21] is used for the shape comparison study. This version contains 9,877 entries, for 9,867 of which the original entry could be retrieved from the PDB [24]. Subsequently, several quality filters for both, small molecule structures and protein crystal structures, are implemented (Fig. 1). Ligand entries are checked for bad valences, inorganic elements, and for their size as function of their number of heavy atoms. Small molecule conformations, as given by the sc-PDB data, are minimized using the MMFF94x force field [25] implemented in MOE 2011.10. The energy difference of the sc-PDB conformation to the minimal energy conformation is calculated and used as a filter criterion to reject compounds in high energy conformations. These criteria lead to a final data set of 7,751 ligands.

Retrieved crystal structure data are analyzed using the Diffraction Precision Index (DPI) [26] calculator developed by Vainio et al. [27]. Entries are kept if the following criteria are fulfilled: resolution ≤ 2.5 Å, $R_{free} \leq 0.3$, and DPI ≤ 0.5 Å. The intersection between the retained 7,258 protein structures and the 7,751 ligand entries yields the final 5,755 protein-ligand pairs.

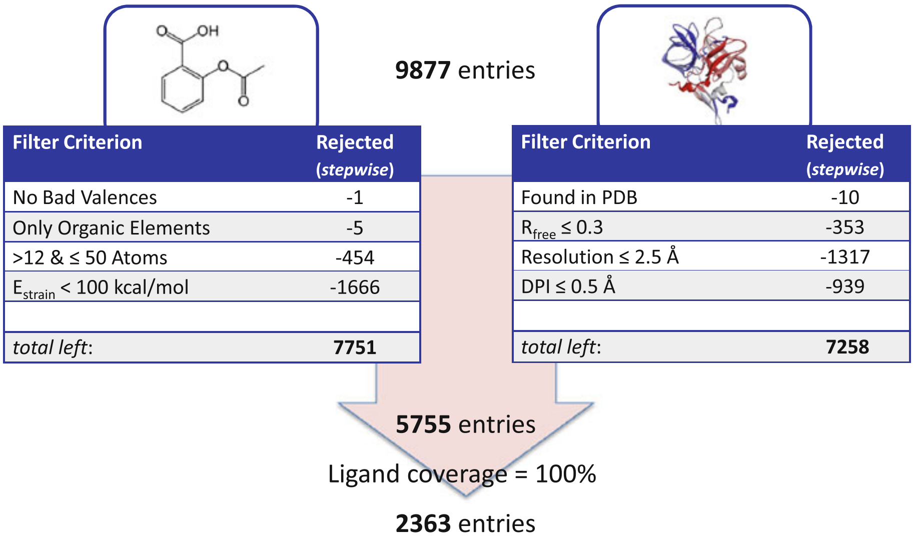
Subsequently, DoGSite [22] is used to identify binding pockets of each ligand. Ligands may be only partially contained in the predicted pocket with parts of it reaching into the solvent or into a neighboring pocket. As such examples bear a potential bias for the comparison between ligand and pocket shape, only pockets with a ligand coverage of 100 % are considered for this analysis, i.e. the entire ligand occupies only one pocket. In total, 2,363 pocket-ligand pairs fulfill this criterion and are kept for the following analyses.

Bioactivity data

Information on experimental ligand bioactivity is retrieved from the PDBBind database [28]. Only entries annotated as “high quality” are considered for this analysis. Entries are matched with data points of the ligand/pocket sc-PDB data set by using the PDB identifier. Overall, 355 experimental values could be matched to ligand/pocket pairs. The data are categorized into three activity classes: highly active: pK_i or $pK_d \geq 8.5$ (83 entries), moderately active: pK_i or $pK_d \geq 6.5$ and < 8.5 (120 entries), and weakly active: pK_i or $pK_d < 6.5$ (152 entries).

Virtual screening data set

Version 13 of the ChEMBL database [29] is used as basis for the docking performance experiments. Data points are



Filter Criterion	Rejected (*stepwise*)
No Bad Valences	-1
Only Organic Elements	-5
>12 & ≤ 50 Atoms	-454
E_{strain} < 100 kcal/mol	-1666
total left:	**7751**

Filter Criterion	Rejected (*stepwise*)
Found in PDB	-10
R_{free} ≤ 0.3	-353
Resolution ≤ 2.5 Å	-1317
DPI ≤ 0.5 Å	-939
total left:	**7258**

Fig. 1 Overview of the sc-PDB data preparation

selected according to the following criteria: all compounds with annotated assay data in IC_{50}, EC_{50}, K_i, or K_d, as standardized by ChEMBL curators, are selected. In order to keep only small molecules, compounds with a molecular weight ≥ 1,000 Da are filtered out. Entries are kept if a curator of the ChEMBL database annotated the observed effect in a particular assay as directly related to an interaction with a particular molecularly defined target (confidence level 9). In case of obvious errors in their annotated bioactivity data, data points are manually removed. Omega 2.4.6 [30] is used to calculate a minimal energy conformer for each molecule. Compounds are rejected from the data set if stereo-chemistry is not properly annotated. In total, 123,539 data points are used for the following analysis. Data points are considered "active" for the analyzed target if a bioactivity below 3.2 μM had been reported. If their annotated bioactivity was above 10 μM or had not been reported with an activity against the target of interest, they were considered "inactive". Note that it cannot be presumed that all of these "inactives" are true negatives. Once tested against the respective target, it is possible that a wrongly assumed inactive may defect to the active set impacting the perceived screening results. The docking experiments are performed on four targets (see Table 1), with the intent to broadly sample the binding pocket NPR shape space (Fig. 11). Additionally, the selected targets are required to hold a sufficient number of measured data points in the extracted ChEMBL data.

Material and methods

Pocket calculation

Potential pockets—solvent exposed and buried ones—are detected on the protein surface using DoGSite [22]. The method maps the protein atoms onto a 3D grid with a default spacing of 0.4 Å. Grid points are labeled as occupied if they are covered by a protein atom; otherwise as free. A difference of Gaussians (DoG) filter is applied to find cavities on the protein surface. Grid points with calculated DoG values below a defined threshold are clustered to subpockets. Subsequently, neighboring subpockets can be merged to pockets. Per default, only pockets and subpockets with volumes larger than 100 and 50 Å^3, respectively, are considered. In this work, we lowered this threshold to 20 Å^3 (subpockets 10 Å^3, respectively) to be included in the pocket shape analysis; the number of pockets per protein was limited to 100.

Note that ligand location is not taken into account for pocket detection. Nevertheless, a ligand can be provided to select the pocket of interest from the set of predicted pockets. In this case, the overlap between ligand and pocket can be calculated, based on the volume overlap of ligand atoms and pocket grid points, resulting in a value between zero and 100 % for pocket and ligand coverage, respectively.

Since subpockets provide a more fine-granular description of the binding site, the experiments in this study are performed on the subpocket level. To simplify the nomenclature throughout this study, the term pocket has been used to signify subpockets, except when both are described in the same context.

Descriptor calculation

Several shape and physicochemical features can be calculated for the detected pockets [23]. In this study, we focus on shape and size, i.e. volume, ellipsoidal shape, and moments of inertia. The discrete pocket volume is calculated by multiplying the number of pocket grid points with the cubic grid spacing. Pocket shape can be represented by ellipsoidal main axes and principal moments of inertia (PMIs). For both representations, pocket grid points are used and weighted with 1 in the calculation. The ellipsoidal pocket description is computed by determining the covariance matrix over all pocket grid points and identifying the corresponding eigenvalues and -vectors. Similarly, PMIs are calculated by a diagonalization of the moment of inertia tensor.

For each ligand the bioactive conformation as reported in the sc-PDB data is chosen. In order to treat protein pockets and corresponding ligands as similar as possible, the atomic weight of each ligand heavy atom is likewise replaced with 1 in the calculation of the inertial moments. Hence, ellipsoids and moments of inertia are calculated based on all ligand heavy atom centers. Hydrogen atoms are not taken into account.

Table 1 Protein crystal structures selected for the virtual screening experiments and their respective active and inactive counts

Target	PDB	Actives (% total)	Inactives
Mitogen-activated protein kinase 14 (p38)	3hrb	1,025 (0.83)	122,514
Androgen receptor (AR)	3g0w	423 (0.34)	123,116
Vascular endothelial growth factor II (VEGFR2)	2qu6	1,328 (1.07)	122,211
Proto-oncogene serine/threonine-protein kinase (PIM1)	3cy3	144 (0.12)	123,395

Normalized principal moment of inertia ratios (NPRs)

Starting from a 3D structure described by either pocket grid points or molecule atoms, the three PMIs are computed. The NPR descriptor is then calculated by sorting the resulting values I_{11}, I_{22}, and I_{33} in ascending order and dividing the two smaller values by the largest. This causes a normalization that eliminates size information and represents one of the major differences in comparison to other molecular shape descriptors that often intrinsically include a description of size. The NPR descriptor can be visualized in a finite, triangular shape space whose corners are representing the three geometrical objects rod, disc, and sphere (Fig. 2). To directly compare the shape character of ligands and pockets, NPR descriptors can be translated into a ternary system describing globularity, disc-, and rod-likeness. Sphericity can be expressed by $(npr_1 + npr_2 - 1)$. Likewise, rod-likeness can be defined as $(npr_2 - npr_1)$ and disc-likeness as $(2 - 2 * npr_2)$, respectively.

The triangular NPR descriptor space is isosceles not equilateral, i.e. only two of the sides are of equal length. For the calculation of distances in this space it is therefore required to transform the descriptor space into an equilateral triangle. This is achieved by an a priori translation of npr_2 to $npr'_2 = \sqrt{3} * npr_2 + (1 - \sqrt{3})$. The pairwise shape distance between two objects A and B is then calculated by the Euclidean distance between the two points $(npr_{1,A'}\ npr'_{2,A})$ and $(npr_{1,B'}\ npr'_{2,B})$. Due to the underlying finite triangular space, the maximal distance between two points is limited to a value of one.

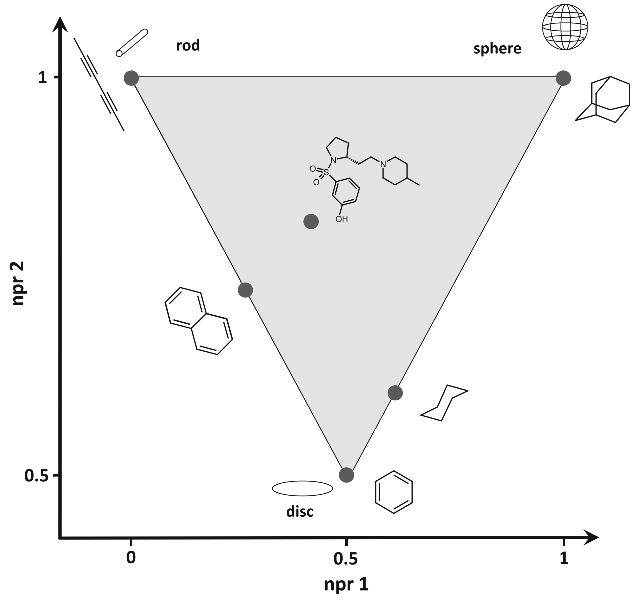


Fig. 2 Overview of the triangular NPR shape space

Screening experiments

Glide HTVS docking is performed using the Maestro software 9.3 of Schroedinger [31]. All protein structures are treated with the default Glide 5.8 settings. Water molecules are removed and hydrogen bonding networks are optimized before grid generation of the receptor by using the Protein Preparation wizard [32]. The binding site is defined by the co-crystallized ligand. Finally, prior to docking, all small molecules from the extracted ChEMBL data are prepared using LigPrep 2.5 and screened with Glide 5.8. Compounds are sorted by ascending docking score, and the best scoring pose is kept for further enrichment analysis. If a compound could not be docked, it is appended at the end of the sorted compound list.

Results and discussion

Pocket shape distribution

In a first analysis, pocket and ligand shapes are directly compared with respect to their distributions in NPR shape space. To do so, the pocket space has previously been restricted to pockets binding a known ligand and exhibiting a volume larger than the predefined cut-off. Due to the size independence of the NPR descriptor, shapes of all kinds of small molecules and binding pockets can be directly compared to each other. The distribution of pocket and corresponding ligands from the 2,363 sc-PDB entries is illustrated in Fig. 3.

Descriptor values for binding pockets do not occupy the western axis of the NPR shape space. While small molecule conformations can be truly flat (e.g., a ligand with conjugated aromatic system), this cannot hold for binding pockets, as this contradicts the definition of a binding site spanning a volume large enough to hold a ligand. The majority of data points of both distributions are located in the upper left corner of the NPR triangle, representing rod-like shapes. The ligand shape distribution of the sc-PDB data set exhibits a characteristic similar to what has been observed for various small molecule data sets of different origins [16]. Similarly, here, no truly globular pockets are observed. While it is difficult to fully explain this behavior, one can speculate about potential contributing factors. First, protein binding pockets and their natural ligands co-evolve. In this context, spherical symmetry of both—pocket and ligand—is a very expensive property to preserve without obvious advantages. Second, truly spherical binding pockets have to be closed on all sides and would need to provide a sort of entry mechanism for their ligands. Such mechanisms are known, for example in the HIV-1 protease, but not common. Third, enzymes in particular

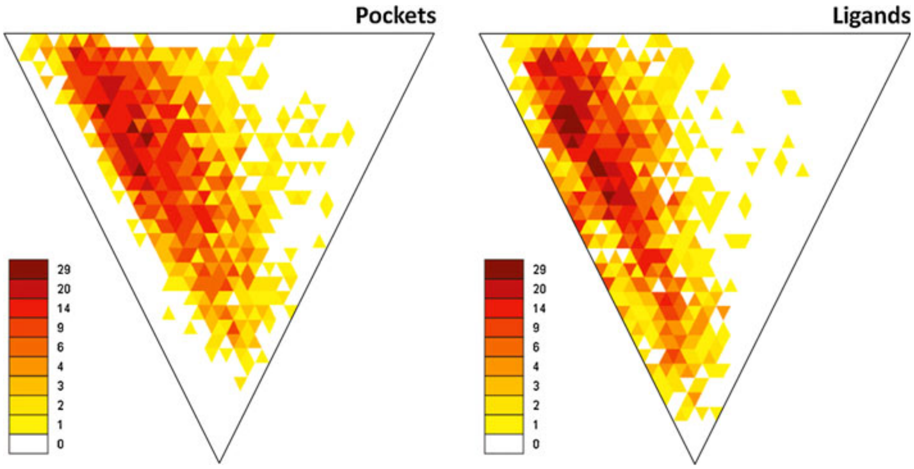


Fig. 3 Shape distribution in NPR space of binding pockets and their respective ligands. *Color code* shows the maximal number of data points per cell

need to incorporate not only the substrate but any reaction partner, transition state and product(s). The lack of symmetry across these requirements further decreases the likelihood for globular binding sites.

The distributions show that binding pockets tend to be slightly more globular than their ligands (Fig. 3). For a more detailed analysis, we investigate the shape difference between corresponding pocket and ligand pairs. Spherical difference, e.g, is calculated as ligand sphericity minus pocket sphericity in percent (for details see methods section). Figure 4 illustrates the pairwise sphere, rod, and disc differences for pairs of ligands and pockets. A difference of zero indicates that pocket and ligand have the same shape character. The distributions within the rod- and disc-likeness difference plots are shifted to the right (Mean (standard error of the mean (SEM)): 0.05 (0.001), and 0.03 (0.001), respectively), while the sphericity peak is shifted to the left (Mean (SEM): −0.08 (0.002). These differences underline the slightly more globular shape character of pockets also on the pairwise level. Only in rare cases, large jumps in NPR space, e.g., rod-like pocket with a discoid ligand, are observed, i.e., distances larger than 0.4 are found in 11.38, 1.18, and 10.16 % of the cases for rod-likeness, sphericity, and disc-likeness, respectively.

In a subsequent study, the full set of detectable pockets is investigated (restricted to a maximum of 100 pockets per structure). For this purpose, the shape of all predicted pockets—ligand-bound or empty—down to a volume of 10 Å^3 is calculated. In contrast to the previously shown ligand-bound pocket shape distribution (Fig. 3), a nearly complete coverage of the NPR shape space can be observed (Fig. 5, left). Interestingly, contrarily to the analysis of occupied pockets, here, the shape distribution center is shifted towards the globular region of the shape space. Furthermore, amongst all detected pockets of a protein, small pockets are observed significantly more often than larger ones (Fig. 5, right).

In order to analyze the dependence of pocket size and shape, the same analysis is repeated for volume ranges encoded by 20 Å^3 bins. The results demonstrate the loss of globularity with increasing pocket volume (Fig. 6). Up to volumes around 100 Å^3 the distribution is dominated by globular pocket shapes. Larger pockets tend to possess a more rod-like shape, expressed by the shift of the shape distribution center towards the upper left corner of the NPR triangle.

The dominance of sphere-like pockets has been described previously in the work of Sonavane and Chakrabarti [18]. Furthermore, Pérot et al. found in a case study of 56 Astex complexes that small pockets are the least compact ones and tend to be rougher, more spherical, and more polar [20]. Two reasons may contribute to this observation: First, frequently encountered ligands of that size are in fact spherical (Na^+, K^+, NH_4^+, Cl^-, SO_4^-) or pseudo-spherical (H_2O without orientation preference). Second, unoccupied binding pockets represent empty space in the protein that is energetically unfavourable. As the sphere has the smallest surface area for a given volume, this shape should be adopted in order to minimize physically unfavourable interactions.

When we split the ligand shape distribution by molecular weight in bins of 100 Da, up to a maximum of 800 Da, these ligand distributions do not exhibit a similar behavior than what has been perceived for protein pockets (data not shown). If any, a slight trend of medium sized ligands towards a more globular shape can be observed.

When comparing ligand and detected pocket shapes, two aspects have to be considered: a) the noise introduced by the pocket boundary definition in the pocket prediction step, and b) the degree to which the analyzed pockets are filled by their ligands. As described in the data preparation section, we have restricted the sc-PDB pocket/ligand shape analysis to pairs for which the ligand is completely contained in the predicted pocket and neglect cases in which

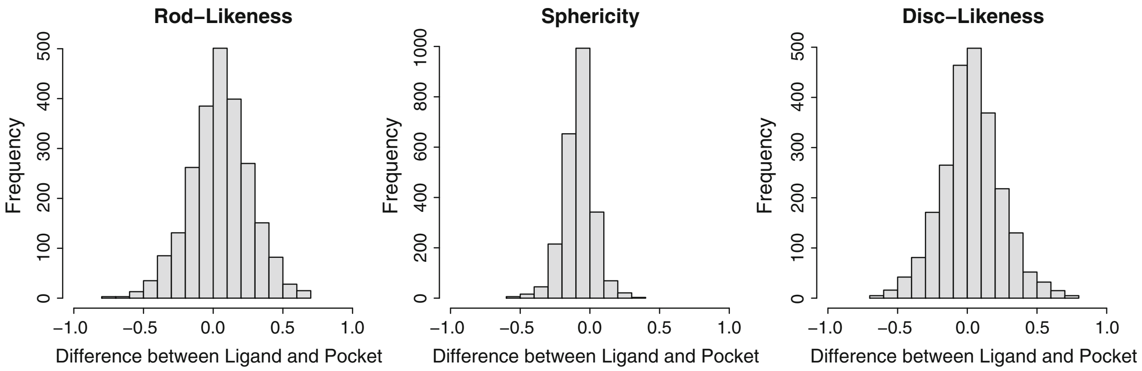


Fig. 4 Pairwise pocket/ligand distance distributions for the three parameters: rod-likeness, sphericity, and disc-likeness, respectivly. Difference values are calculated by subtracting pocket values from ligand values

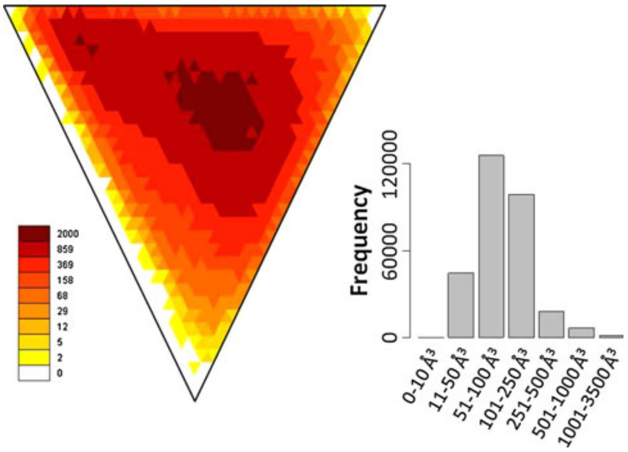


Fig. 5 NPR shape distribution of all pockets detected by DoGSite (with disabled 100 Å^3 volume threshold). *Color code* shows the maximal number of data points per cell. Histogram depicts the volume distribution of these pockets

parts of the molecule are located outside the identified pocket. However, the analysis, so far, has not taken pocket coverage into account. Thus, molecules can be much smaller than their complementary binding pocket and occupy only a small part of it. Therefore, the shape behavior in dependence of pocket coverage, split into 5 bins of equal size, is analyzed. In this experiment, pocket and subpocket based performance is regarded explicitly. When considering pockets, a clear peak is detected at the 20–40 % coverage bin, indicating that a "golden ratio" between ligand and pocket volume is prevalent in the data set (Fig. 7).

This confirms the findings by Kahraman et al. [4] on 100 binding sites of 9 frequent ligands that generally only one third of the pocket is covered by the ligand. Also, Pérot et al. [20] stated that pockets generally tend to bind smaller ligands. Nevertheless, analyzing the subpocket-based behavior, coverages between 40 and 60 % are preferentially observed. The tighter pocket boundary definition of subpockets allows for a more fine-granular analysis and indicates a higher shape complementary than previously found.

The analysis of pairwise distances between ligand and pocket shapes as a function of pocket coverage reveals that ligand/pocket pairs are more similar in shape at higher coverage values (Fig. 8a). For pairs with coverages between 80 and 100 %, the distance in NPR space is on average 0.15, compared to a baseline distance of 0.24 derived from all ligand shapes against all pocket shapes.

In addition to shape complementarity, the ligand's location and orientation in the pocket is analyzed. In Fig. 8b the centers-of-mass distances of pockets and ligands are compared with respect to coverage. Understandably, the distance of the centers-of-mass diminishes with increasing pocket coverage.

The mean distance for poorly covered pockets lies at 7.13 Å, while the number decreases to 0.75 Å for perfectly overlaying volumes. To quantify the orientation of ligand and pocket to each other, the angle deviation between the respective largest ellipsoidal main axes is calculated. Again, angle deviations are analyzed with respect to pocket coverage, yielding the conclusion that ellipsoidal alignment increases with pocket coverage. Figure 8c illustrates that the median angle of pockets covered to 40–60 % lies at 28.79° and drops to 9.66° for almost completely covered pockets. For all three discussed parameters, a clear gliding correlation with increasing pocket coverage is observed.

Throughout these studies, a few outliers have been observed, e.g., large center-of-mass distances or angle deviations. These cases may originate from the pocket prediction step, as large proteins, multimers, or generally the large variety in binding site space make pocket detection with accurate boundary definitions difficult.

Ligand binding demands a decent amount of congruence between the binding partners. Within this analysis, we found that this required shape complementary can generally be quantified by a subpocket coverage above 50 %, a shape distance in NPR space lower than 0.2, a center-of-

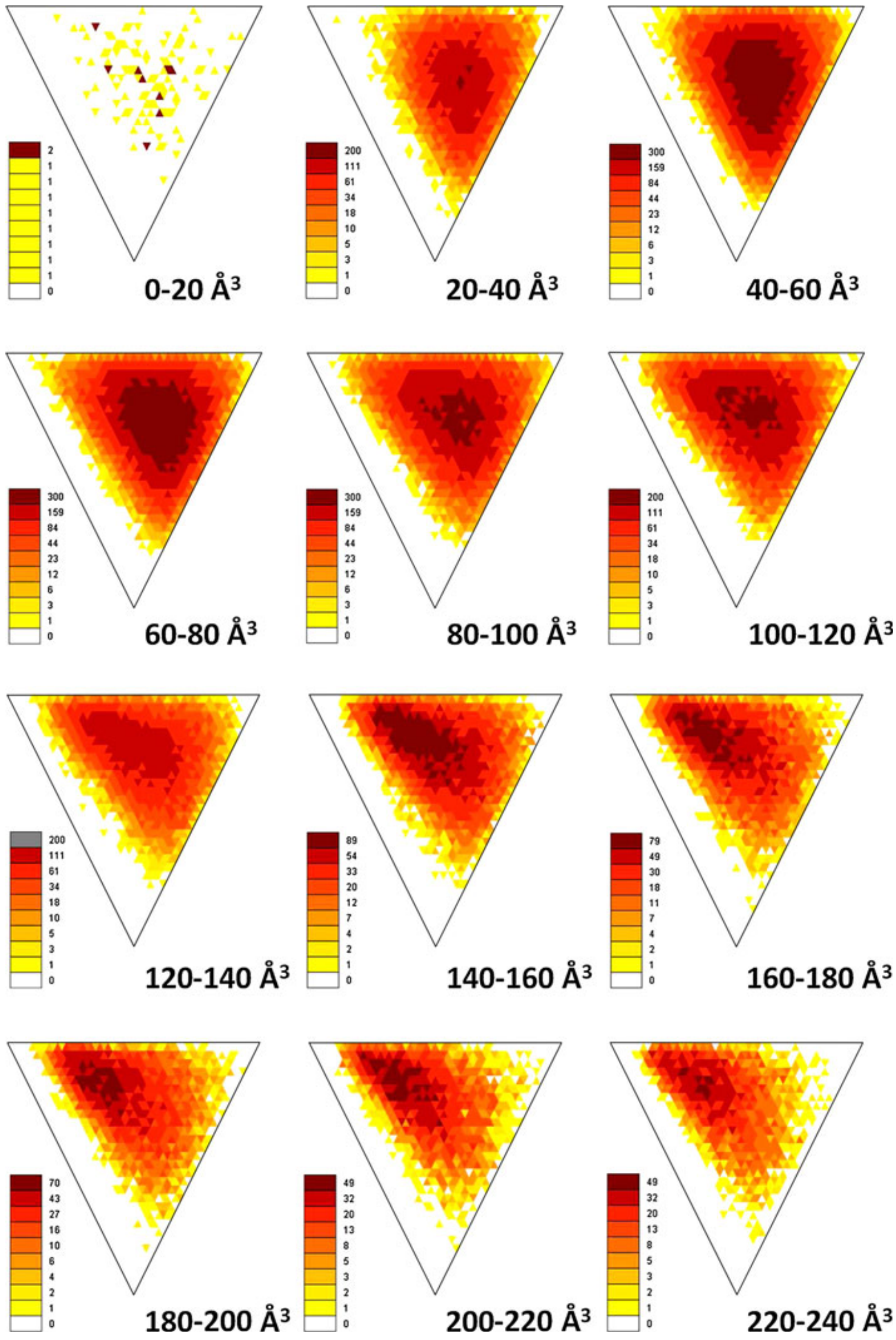


Fig. 6 NPR shape distribution by pocket volume bins. Color code shows the maximal number of data points per cell

mass distance lower than 2.4 Å and a main axis angle deviation below 29°.

Bioactivity, ligand efficiency, and pocket parameters

A popular assumption in drug design is that binding affinity is driven by shape complementarity of the ligand with the protein [7], as the geometrical fit is a prerequisite of binding. Matching pharmacophoric features within appropriate distance and correct directionality, as well as angles of established hydrogen bonds, are other important driving factors for molecular recognition. The relative importance of electrostatic complementarity *vs*. shape complementarity remains an open question. To understand to which degree shape complementarity alone is responsible for activity, the correlation between bioactivity, as extracted from the PDBbind database for 355 examples, and different pocket parameters is assessed in this study (Table 2). While no strong correlations can be identified, the generally expected trends are confirmed. Pocket coverage exhibits a moderate

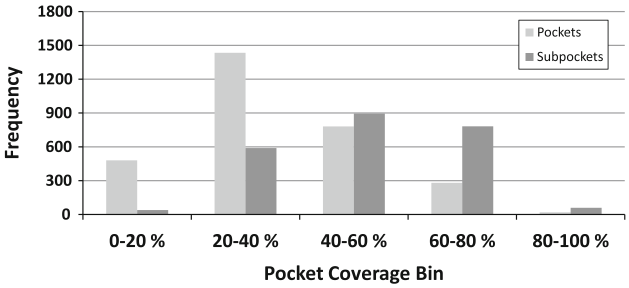


Fig. 7 Pocket coverage distribution of pockets and subpockets in sc-PDB data set

correlation to activity. Furthermore, distance in NPR shape space, distance between the centers-of-mass of ligand and pocket, as well as the angle between the first principal axes of ligand and pocket ellipsoid show moderate negative correlation to activity.

These trends become more pronounced when pocket/ligand pairs are binned into three categories based on their bioactivity. Significant differences between the distributions for the previously mentioned pocket parameters—pocket coverage, NPR distance, centers-of-mass distance, and ellipsoid angle deviation—can be observed (Fig. 9). The trends within these data points mirror the respectively found correlations between the classifiers.

Work from Kuntz [33] and Reynolds [34] assesses the maximal achievable binding affinity and efficiency of a small molecule in dependence of its size. Their results show that maximal binding efficiency flattens off for molecules of large size which they attribute to enthalpic, entropic, and geometric reasons. Similarly, we analyzed whether there might be a possible limitation originating from the size of the binding pocket. Such information could provide a better understanding regarding what binding efficiency could be maximally expected for a given pocket, and hence be an additional parameter for target druggability assessments. Here, we use NBEI (=*pKi/Number of Heavy Atoms*) [35] as ligand efficiency index. Figure 10a shows for the 355 extracted PDBbind ligands a similar decline of maximal efficiency with increasing size of the molecule, as has been found by Kuntz and Reynolds. Figure 10b highlights the distributions of NBEI with respect to the corresponding binding pocket volume bins as an alternative size measure. A similar decrease of efficiency with increasing volume of the identified pocket can be seen. Above a volume of 1,200 Å^3, maximal NBEIs observed are significantly lower than in smaller volume bins. Additionally, the data suggest that targets containing a binding pocket with a volume between 300 and 700 Å^3 have a higher probability for the identification of a highly efficient binder. Focusing on volumes, Liang et al. [17] describe a similar correlation between ligand and pocket dimensions, predominately for pocket volumes with less than 700 Å^3. Generally, value comparisons between studies are difficult due to differences in the computation. Furthermore, even when considering several structures of the same target bound to different ligands, predicted pockets and their properties vary. In a previous study on druggable pockets [23], the volumes of 40 p38 kinase ligand binding sites in different activitation states, predicted with DoGSite, were found to span a volume range from 450 Å^3 to almost 1,800 Å^3 (pocket, not subpocket values). This clearly shows the impact of the ligand on pocket shape and volume.

Interestingly, no correlation (R = −0.01) between NBEI and pocket coverage can be observed, while there exists a moderate correlation between ligand size as number of heavy atoms and pocket coverage (R = 0.41). Incorporating the information that activity has a moderate

Table 2 Pearson correlation coefficient of activity versus other pocket parameters

Correlation of activity versus	R (Pearson)
Pocket coverage	0.39
Distance NPR ligand/pocket	−0.32
Distance centers-of-mass ligand/pocket	−0.34
Angle between first principal ellipsoid axes	−0.29

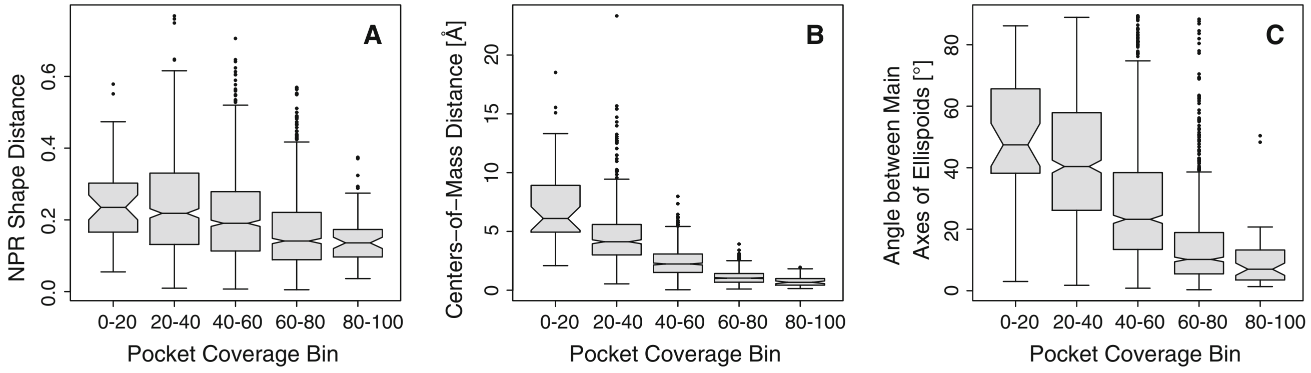


Fig. 8 Coverage-binned distributions of **a** pairwise ligand and pocket euclidean distance in NPR shape space, **b** distances between the centers-of-mass, **c** angles between the first main ellipsoidal axes of pocket and ligand ellipsoid

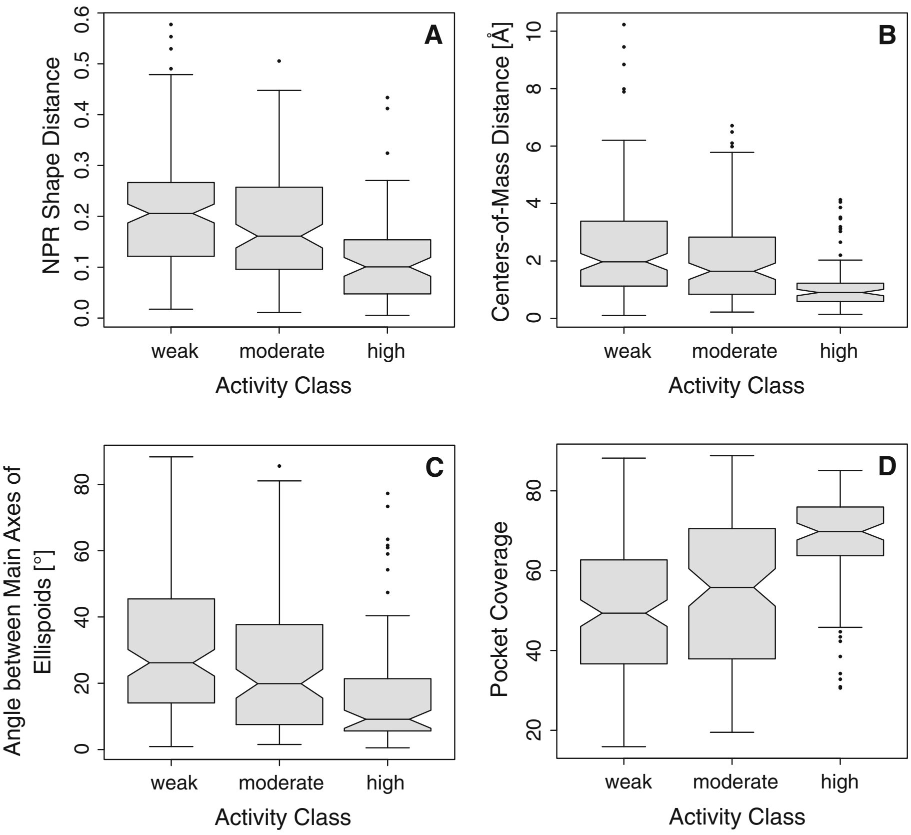


Fig. 9 Distributions of NPR shape distances (**a**), centers-of-mass distances (**b**), angles between ellipsoidal main axes (**c**) and pocket coverage (**d**) of weakly, moderately, and highly active compounds and their respective pockets

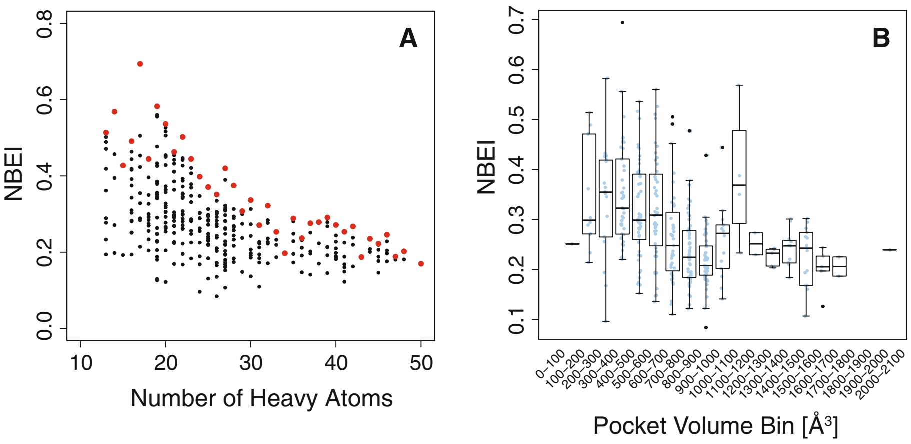


Fig. 10 **a** NBEI *vs. Number of Heavy Atoms*. Maximal values of NBEI for each *Number of Heavy Atoms* value are highlighted in *red*. **b** NBEI distributions stratified by binding pocket volume bins

correlation to pocket coverage, it can be concluded that with rising pocket coverage, activity is enhanced as binding pocket and ligand complementarity increases. This seems to be more prevalent for larger ligands, but as binding efficiency is determined as the ratio between activity and size, this has no effect on efficiency.

Protein/ligand shape and size comparison in virtual screening

The capability of different parameters retrieved from perceived binding pockets and corresponding co-crystallized ligands to identify active molecules is analyzed using a selected data set of 123,539 compounds from the ChEMBL database. Four molecular targets [PIM1 (PDB code: 3CY3), AR (PDB code: 3G0W), p38 (PDB code: 3HRB), VEGFR2 (PDB code: 2QU6)] have been chosen to broadly sample the binding pocket NPR shape space (Fig. 11).

We compare the effects of five different parameters to each other regarding early enrichment in virtual screening: Euclidean distance in NPR shape space between ChEMBL compounds and co-crystallized ligand or binding pocket, respectively, absolute volume difference between ChEMBL compounds and co-crystallized ligand, absolute

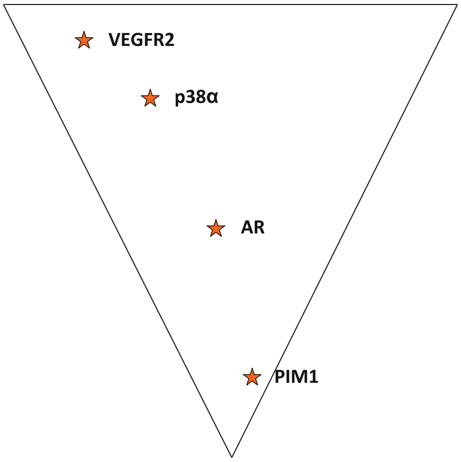


Fig. 11 Overview on NPR shape space location of the selected binding pockets of the selected targets

difference in number of heavy atoms between co-crystallized ligand and ChEMBL compounds, and absolute compound to pocket volume ratio distance from a predefined gold standard value. The absolute volume ratio distance is calculated as $|0.4 - compound\ volume/pocket\ volume|$. The value 0.4 has been chosen based on the mean values found during the previous analysis of pocket and subpocket coverage. To assess active retrieval performance in comparison to random, the early enrichment factor (EEF) is determined as $EEF_{x\%} = \%Actives_{x\%}/\ \%Data_{x\%}$. Figure 12 illustrates the respective enrichment results at x = 1, 2.5, 5, 7.5, and 10 % of the data. Corresponding ROC curves and AUCs can be found in the supplemental material (Figure S1). For the evaluation of active retrieval performance, we favor EEFs over AUCs, as in most experiments only a small percentage of the full dataset will be selected and tested experimentally. The performance on target p38 is close to random; only for the 1 and 2.5 % bins a few parameters, e.g., number of heavy atoms and NPR pocket distance, show higher EEF values than one. In the example of AR, all parameters, especially those considering volume differences, show a better enrichment than what could be expected by chance. For PIM1, these two volume dependent filters are the only parameters showing a good enrichment better than random. In contrast, for VEGFR2, NPR distance to the co-crystallized ligand clearly outperforms all other studied parameters.

The generally good performance of the ligand-based filters suggests that shape and size similarity to a reference ligand, if available, might be a more valuable parameter to optimize for than a high similarity with the binding pocket, especially considering NPR shape. Note, that the pocket-based NPR shape filter performs better for targets with binding site shapes closer to the center of ligand shape distribution, e.g., for AR (see Fig. 11). The filter has a higher probability to fail in cases where the binding site shape has a large distance to this center of distribution, e.g., in the case of PIM1 whose binding pocket shape is located close to the disc-like corner. While shape similarity based on distance in NPR shape space is superior in p38 and VEGFR2, results from PIM1 and AR show a better performance when using the difference in volume instead. Considering volume or volume ratio distance, the pocket parameter performance is similar to the ligand-based filter. This is due to the fact that this volume ratio, although solely based on the pocket volume, describes the difference in volume between an on average fitting reference molecule and the compound of interest. Thus, this provides very valuable information in cases where no reference molecule is known.

Averaged enrichment factors over all analyzed early enrichment bins (1, 2.5, 5, 7.5, and 10 %) and targets showed similar results for absolute volume ratio distance of compound and binding pocket, NPR distance and volume difference based on comparison with the co-crystallized ligand with a factor of 1.7 over random performance. Difference in the number of heavy atoms and NPR distance between binding pocket and screened compound perform worse (average EEF of 1.3 and 1.1, respectively). Although utilizing the selected parameters as ranking scheme for compounds results in a better performance than what could be expected by chance, it is clear that the achieved performance is not sufficient for practical purposes; too many of the known actives would be missed during experimental testing. As a comparison, docking experiments against crystal structures of the four chosen targets are performed using the Glide HTVS docking engine [31]. By integrating an elaborated scoring function, an increase in signal over the previous experiment is to be expected. In general, the introduction of Glide improves the achieved average EEF from the average value of 1.5 of the discussed parameters to 7.4 over all analyzed early enrichment bins (1, 2.5, 5, 7.5, and 10 %) and targets (Table 3).

Three of the selected targets, namely p38, AR, and VEGFR2, are part of the Directory of useful Decoys (DUD) [36] and have been analyzed in a benchmark study by Cross et al. [37]. A mean enrichment factor for all DUD targets between 13.3 (at 0.5 %) and 3.8 (at 10.0 %) was reported by Cross et al., which closely resemble the mean EEFs between 13.5 (at 0.5 %) and 4.1 (at 10.0 %) for the four targets analyzed in this study.

However, placing and scoring several thousands of compounds with high precision is a very time-consuming process. The aim of this analysis is therefore to investigate whether information on binding pocket and co-crystallized ligand can be used to pre-filter the data set. Ideally, a filter increases early retrieval rates of actives by removing compounds that do not fulfill the requirements of the molecular target. Generally, decreasing the number of

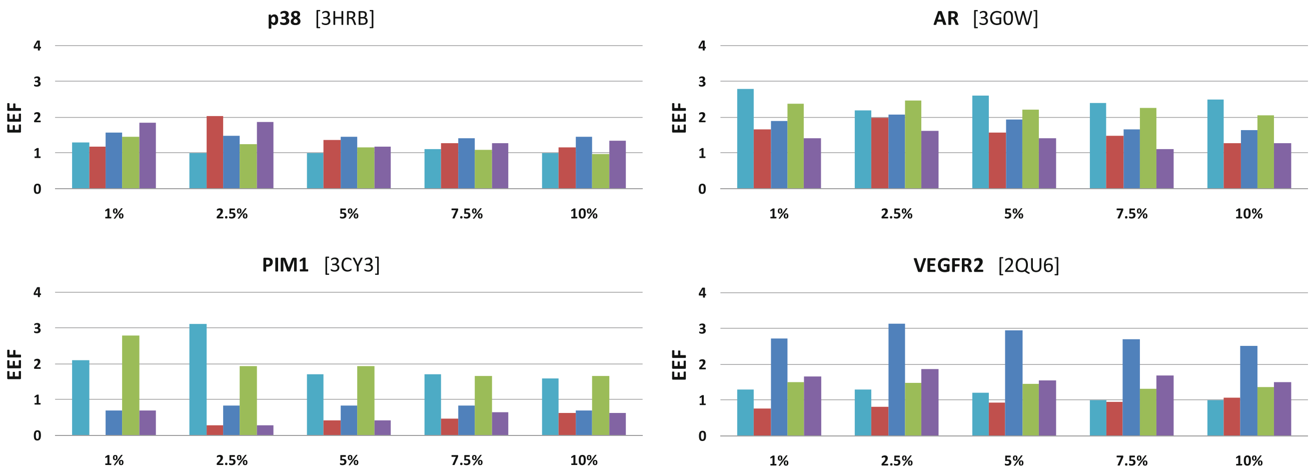


Fig. 12 Early Enrichment (%Actives against %Data screened) for p38 (3HRB), Androgen receptor (3G0W), PIM1 (3CY3) and VEGFR2 (2QU6). Volume ratio difference between pocket and screened compound (turquoise). NPR distance between pocket (*red*)/reference ligand (*blue*) and screened compound. Volume difference between reference ligand (*green*) and screened compound. Difference in *Number of Heavy Atoms* between reference ligand and screened compound (*purple*)

Table 3 Early enrichment factors (EEF) by glide for the selected targets

Target	$EEF_{1\%}$	$EEF_{2.5\%}$	$EEF_{5\%}$	$EEF_{7.5\%}$	$EEF_{10\%}$
p38 [3HRB]	19.9	10.6	6.6	4.9	4
AR [3G0W]	19.6	11.4	7.7	5.6	4.5
VEGFR2 [2QU6]	8.9	5.6	3.8	3.1	2.8
PIM1 [3CY3]	5.6	6.7	6.0	5.4	5.0

compounds to screen while achieving constant early enrichment rates is equally important, as this speeds up the docking process and requires less computational resources. Exemplarily, we investigate a reduction of the data set by one third based on the chosen parameters. When removing 33 % of the compounds based on the chosen parameters (Fig. 13), early enrichment factors improve for PIM1. The performance slightly decreases for the two targets, AR and VEGFR2, when considering NPR shape filters. In contrast, considering volume and molecular size-based filters, the performance in AR remains constant, while it slightly increases for VEGFR2. For p38, enrichment factors decrease for all tested parameters, particularly for the NPR based filters. As null hypothesis, we carry out a random filtering that is removing 33 % of the data in 500 independent runs. The resulting $\Delta EEFs$ in Fig. 13 show that all filters perform better than random, except for target p38, in which the NPR distance filter based on a reference ligand does not show an improvement over a random removal of compounds prior to the virtual screening experiment. Corresponding ROC curves and AUCs can be found in the supplemental material (Figure S2).

To summarize, the results show that the introduction of filters based on molecular or pocket shape and size properties can aid in decreasing the number of compounds to screen without distinctly diminishing performance in active retrieval. The results furthermore suggest that a molecular size-based pre-filter might be more effective than a shape-based pre-filter in structure-based virtual screening. It is a known effect that docking scores can be correlated with molecular size [38]. We cannot exclude that this might be the reason for the observed effect in the performed study. The tested parameters based on ligand information appear to be more valuable for the pre-filter application. Nevertheless, a pocket-based filter can be very useful in screening scenarios, in which no bound reference ligand is known beforehand. The calculated compound to pocket volume ratio distance, e.g., mimics the volume distance of a compound from a potentially binding ligand under the assumption that the most prominent volume ratio between bound ligand and pocket lies at 0.4. While this value has been derived from the previous analysis, it holds for the four studied targets, in which pocket and ligand volume ratios evaluate exactly to this value. Also, while we describe observations for different parameters, it is unclear whether these results can be extrapolated to other data sets, targets, pocket prediction algorithms or docking programs. We are nevertheless convinced that this study gives an idea on how filters could be usefully integrated into a virtual screening process.

Conclusion

In this study, a comparison between shapes of protein binding pockets and their corresponding ligands using the NPR shape description is presented. The experiments within this study are performed on predicted subpockets

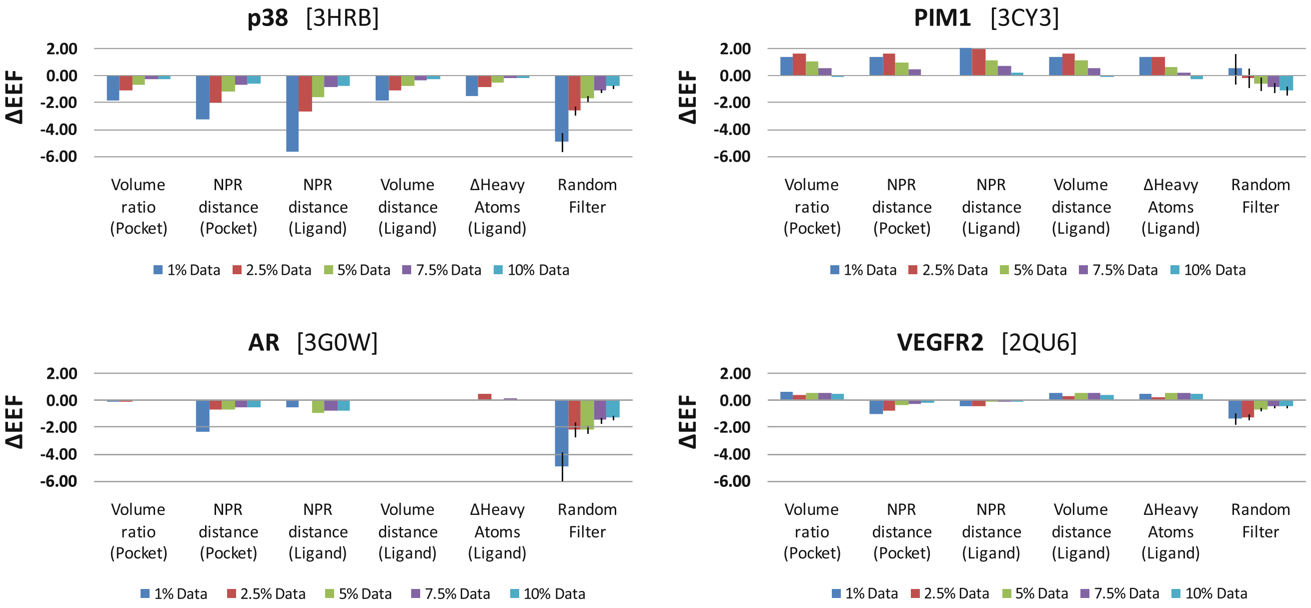


Fig. 13 Differences for p38 (3HRB), Androgen receptor (3G0W), PIM1 (3CY3), and VEGFR2 (2QU6), respectively, in early enrichment factors to the original Glide enrichment factor (Table 3) when filtering 33 % of the data based on the given parameters. For comparison, results of a random 33 % filter are presented. *Bars* represent the mean, lines the standard deviation of results from 500 independent runs

(although the term pocket is used throughout the study), which provide a finer description of a cavity. Similarly to what has been reported for small molecule data sets of various origins, we observe an absence of truly spherical ligand and binding pockets in our stringently selected data set of 2,363 pocket/ligand pairs from the sc-PDB. Contrarily, globular pockets are prevalent when considering all detectable pockets within protein structures, particularly for pockets of small volume. Furthermore, when considering pockets (not subpockets), it could be seen that on average the ligand covers one third of the pocket. This is in accordance with findings from Kahraman et al. [4] on a less diverse data set. Subpockets show on average a higher coverage (50 %), underlining the information gain due to the more fine-granular description. Thus, this study broadens the knowledge about the importance of shape complementary for complex formation. A clear decrease of three analyzed parameters with increasing pocket coverage can be observed: The average shape distance is found to be below 0.2 in NPR space, the centers-of-mass distance of the respective ellipsoids is below 2.4 Å, and the angle between the largest ellipsoidal main axes deviates less than 29°. While it is difficult to provide gold standard values, due to different implementations of pocket detection algorithms and ways to compute features thereof, in our opinion, these results provide a very valuable starting point for further research.

With respect to exhibited bioactivity, tested on a PDBBind subset, only a moderate correlation with pocket coverage is observed, which illustrates the importance of other complementing factors. However, when binning active compounds into three classes by their strength of interaction, it was found that high affine binders tend to possess a higher shape complementarity and cover on average more than two thirds of the binding pocket. Also, their corresponding molecular shapes are significantly more strongly aligned. Contrarily, low-affine binders exhibit a larger flexibility in their shape congruence to the binding site. With respect to ligand efficiency, this article discusses to which extent pocket size influences the maximal achievable ligand efficiency. The data suggest that pockets with a volume smaller than 700 Å^3 have an increased probability of fitting a highly efficient binder. Although these findings need to be investigated further and the value is dependent on the pocket identification method used, this provides a very interesting starting point into the investigation of pocket druggability. Clearly, maximal achievable binding efficiency cannot only be related to the size of a small molecule but also to features present in the corresponding binding site.

This article furthermore investigates whether it is possible to use information on binding pocket volume and shape to filter out compounds a priori in a large-scale virtual screening campaign. The hypothesis is that such a step might increase the hit rate of the screen by preferentially removing inactive compounds and simultaneously decreasing the number of compounds to be screened. With regards to the ever-increasing number of compounds available for virtual screening and with the introduction of more sophisticated calculations, the reduction of compound sets is of high interest. As can be seen on the four selected examples, using the provided filters reduces the number of compounds by one third, while achieving mostly stable early enrichment rates. Filters making use of information extracted from known reference molecules show generally a good performance. Nevertheless, using pocket information, especially if no reference ligand is known, enables a new set of filters applicable in structure-based virtual screening.

Acknowledgments We thank Volker Hähnke and Serge Christmann-Franck for constructive discussions and Jeffrey Shaw for performing the Glide docking runs. Andrea Volkamer acknowledges funding from the BMBF (Grant 0315292A) for the pocket analysis project as part of the Biokatalyse2021 cluster. Matthias Wirth thanks Merck Serono S.A. for a PhD fellowship.

References

1. Náray-Szabó GG (1993) J Mol Recognit 6(4):205
2. Jennings A (2011) In: Tari LW (ed) Structure-based drug discovery methods in molecular biology. Springer Protocols, Human Press, pp 235–250
3. Chen K, Kurgan L (2009) PLoS ONE 4(2):e4473
4. Kahraman A, Morris RJ, Laskowski RA, Thornton JM (2007) J Mol Biol 368(1):283
5. Kahraman A, Morris RJ, Laskowski RA, Favia AD, Thornton JM (2010) Proteins 78(5):1120
6. Nicholls A, McGaughey G, Sheridan R, Good A, Warren G, Mathieu M, Muchmore S, Brown S, Grant J, Haigh J, Nevins N, Jain A, Kelley B (2010) J Med Chem 53(10):3862
7. Morris R, Najmanovich R, Kahraman A, Thornton J (2005) Bioinformatics. Oxford, England. 21(10):2347
8. Putta S, Beroza P (2007) Curr Top Med Chem 7(15):1514
9. McGaughey G, Sheridan R, Bayly C, Culberson J, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell W (2007) J Chem Inf Model 47(4):1504
10. Kortagere S, Krasowski M, Ekins S (2009) Trends Pharmacol Sci 30(3):138
11. Rush T, Grant J, Mosyak L, Nicholls A (2005) J Med Chem 48(5):1489
12. Miller MD, Sheridan RP, Kearsley SK (1999) J Med Chem 42(9):1505
13. Ballester P, Richards W (2007) J Comput Chem 28(10):1711
14. Sauer W, Schwarz M (2003) J Chem Inf Comput Sci 43(3):987
15. Akritopoulou-Zanze I, Metz J, Djuric S (2007) Drug Discov Today 12(21–22):948
16. Wirth M, Sauer W (2011) Mol Inf 30:677
17. Liang J, Edelsbrunner H, Woodward C (1998) Protein Sci Publ Protein Soc 7(9):1884
18. Sonavane S, Chakrabarti P (2008) PLoS Comput Biol 4(9):e1000188
19. Weisel M, Kriegl JM, Schneider G (2010) ChemBioChem 11(4):1
20. Pérot S, Sperandio O, Miteva M, Camproux AC, Villoutreix B (2010) Drug Discov Today 15(15–16):656
21. Meslamani J, Rognan D, Kellenberger E (2011) Bioinformatics 27(9):1324
22. Volkamer A, Griewel A, Grombacher T, Rarey M (2010) J Chem Inf Model 50(11):2041
23. Volkamer A, Kuhn D, Grombacher T, Rippmann F, Rarey M (2012) J Chem Inf Model 52(2):360
24. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) Nucleic Acids Res 28(1):235
25. Halgren T (1996) J Comput Chem 17(5–6):490
26. Blow DM (2002) Acta Crystallographica Section D 58(5):792
27. Vainio MJ, Puranen JS, Johnson MS (2009) J Chem Inf Model 49(2):492
28. Wang R, Fang X, Lu Y, Wang S (2004) J Med Chem 47(12):2977
29. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, O-verington JP (2011) Nucleic Acids Res 40(D1):D1100
30. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) J Chem Inf Model 50(4):572
31. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) J Med Chem 47(7):1739
32. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W (2013) J Comput-Aided Mol Des 27(3):221
33. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) Proc Natl Acad Sci USA 96(18):9997
34. Reynolds CH, Tounge BA, Bembenek SD (2008) J Med Chem 51(8):2432
35. Abad-Zapatero C, Perisic O, Wass J, Bento AP, Overington J, Al-Lazikani B, Johnson ME (2010) Drug Discov Today 15(19–20):804
36. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49(23):6789
37. Cross J, Thompson D, Rai B, Baber J, Fan K, Hu Y, Humblet C (2009) J Chem Inf Model 49(6):1455
38. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P (2004) J Chem Inf Comput Sci 44(3):793