JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

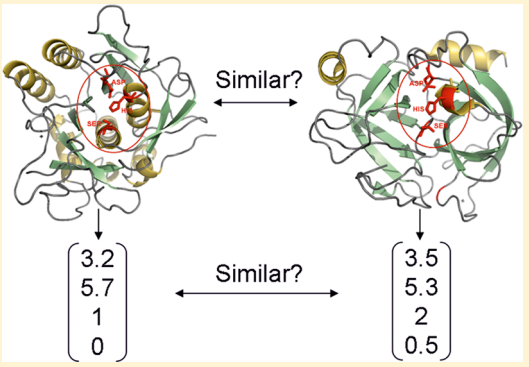
Alignment-Independent Comparison of Binding Sites Based on DrugScore Potential Fields Encoded by 3D Zernike Descriptors

Britta Nisius and Holger Gohlke*

Department of Mathematics and Natural Sciences, Institute of Pharmaceutical and Medicinal Chemistry, Heinrich-Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

Supporting Information

ABSTRACT: Analyzing protein binding sites provides detailed insights into the biological processes proteins are involved in, e.g., into drug−target interactions, and so is of crucial importance in drug discovery. Herein, we present novel alignment-independent binding site descriptors based on DrugScore potential fields. The potential fields are transformed to a set of information-rich descriptors using a series expansion in 3D Zernike polynomials. The resulting Zernike descriptors show a promising performance in detecting similarities among proteins with low pairwise sequence identities that bind identical ligands, as well as within subfamilies of one target class. Furthermore, the Zernike descriptors are robust against structural variations among protein binding sites. Finally, the Zernike descriptors show a high data compression power, and computing similarities between binding sites based on these descriptors is highly efficient. Consequently, the Zernike descriptors are a useful tool for computational binding site analysis, e.g., to predict the function of novel proteins, off-targets for drug candidates, or novel targets for known drugs.



INTRODUCTION

Biological processes are based on interactions between proteins and other molecules, e.g., ligands, other proteins, or nucleic acids. Usually, these interactions occur at defined locations, the binding sites. There are several types of binding sites in proteins, e.g., active sites in enzymes, sites for allosteric regulation, or binding interfaces of protein−protein interactions.[1] Shape complementarity and physicochemical complementarity are key factors for molecular recognition and interactions.[2] Therefore, a binding site's shape, size, and buriedness, as well as its amino acid composition and, hence, its physicochemical properties, are important properties for its characterization. This complexity makes the analysis of binding sites a demanding problem.[3]

Detecting biologically relevant similarities between protein binding sites is a main application of computational binding site analysis[4] because it allows (I) predicting a drug candidate's off-target interactions,[3] (II) predicting new targets for known drugs,[5−7] or (III) designing selective drugs by analyzing (dis)similarities among members of one target family.[8] Likewise, computational approaches for binding site comparison have been applied for protein function-deorphanization[9] assuming that proteins with similar biochemical functions should bind similar ligands and, hence, display similar binding site characteristics.

Various approaches aiming at the computational identification, analysis, and comparison of protein binding sites have been developed in recent years.[10] Unless a cocrystallized ligand readily provides information on the binding site's location, predicting this location is the first step. Approaches aiming at this scan the surface of the protein for pockets or cavities that most likely are binding sites.[11,12] Since many of these methods have been developed in recent years,[11] we decided to solely focus on the analysis of already defined binding sites rather than trying to tackle the problem of pocket detection, too. In order to computationally analyze a binding site, its features have to be transformed to numerical values enabling an efficient analysis. To do so, structure-based approaches for computational binding site analysis typically use simplified representations of the amino acids flanking the binding sites, which are encoded as geometric patterns or numerical fingerprints.[13] For a comparison, most of the methods require the binding sites to be optimally aligned, which is often time-consuming. Furthermore, an incorrect alignment of two binding sites can lead to dramatically underestimated similarity scores.[14] By contrast, only very few alignment-independent representations are available.[14−17] In the present study, we thus aim at developing novel alignment-independent binding site descriptors that can be used for protein binding site comparison as well as for other principal applications in computational binding site analysis, such as druggability prediction.[18]

While many approaches characterize a binding site based on its shape[17,19] or on the amino acids flanking it,[20−22] the molecular recognition between a protein and a ligand is governed by interactions that occur *within* the binding site.

Received: May 24, 2012
Published: August 10, 2012

dx.doi.org/10.1021/ci300244y | J. Chem. Inf. Model. 2012, 52, 2339−2347

Thus, we decided to describe protein binding sites based on their molecular recognition properties encoded in molecular interaction fields (MIFs). MIFs are popular tools for characterizing the ability of a molecule to interact with other molecules.[23] The principal idea of MIFs was introduced by Goodford.[24] Since then, MIFs have frequently been applied for the analysis of ligands, proteins, and protein binding sites. In the latter case, MIFs have been used for analyzing structural differences within large target families, e.g., proteases,[25] protein kinases,[26] or nuclear receptors.[27] Again, most of these approaches require an alignment of the proteins prior to analysis, although a few alignment-independent approaches for the computational analysis of MIFs have been introduced.[15,16] These methods first simplify the MIF by extracting only representative and energetically favorable regions (hot spots). Second, the hot spots are encoded in an alignment-independent way, e.g., by defining pharmacophore patterns[15] or by autocorrelation analysis.[16] In these two-step approaches, extracting the hot spot regions from the MIF is the most crucial, but also the most difficult step.[28]

Thus, we intended to develop an approach allowing an alignment-independent encoding of MIFs that does not require an initial extraction of hot spot regions but rather makes use of the complete information of a MIF. To this end, we use the DrugScore[29] scoring function for the computation of MIFs. The resulting DrugScore potential fields are then transformed to a set of alignment-independent descriptors via a series expansion in 3D Zernike polynomials.[30] So far, 3D Zernike function expansion was mainly used for shape retrieval,[31] i.e., for comparing the shapes of proteins,[32−34] ligands,[35] and protein binding sites.[17] However, a series expansion in Zernike polynomials is not limited to binary 3D objects (where, e.g., 1 signals "in" the molecule and 0 "out"). Rather, using these polynomials any 3D object can be encoded as a set of rotation-invariant descriptors and, hence, also MIFs.

MATERIALS AND METHODS

Characterizing Protein Binding Sites by DrugScore Potential Fields. To characterize binding pockets by MIFs, a cubic grid is placed within the binding site. Then, a probe atom of a particular atom type is placed at every grid point, and the interaction between this probe atom and all amino acids flanking the binding site is quantified by the distance-dependent, knowledge-based pair potentials of the DrugScore[29] scoring function (Figure 1).

We decided to use DrugScore for the MIF computation because it has been shown to be successful in predicting correct binding modes,[36−38] thus demonstrating that DrugScore potential fields correctly characterize molecular recognition properties of protein binding sites. Furthermore, based on the formalism of the DrugScore approach, distance-dependent pair potentials for RNA-ligand[39] and protein−protein[40] interactions have been derived, enabling the computation of MIFs, and subsequently Zernike descriptors, also for these systems.

For all computed DrugScore potential fields, a grid spacing of 0.375 Å was used. By using probe atoms of different atom types, different types of interactions within the binding site can be characterized. In this study, we utilized five probe atoms, each one encoding a particular type of protein−ligand interaction (Table 1).

When computing DrugScore potential fields, the grid box is usually centered on a ligand placed within the protein structure. If DrugScore potential fields are computed for ligand-free protein structures, the dimensions of the grid box can be defined manually, e.g., after aligning the *apo* structure to a *holo* structure of the same protein. Alternatively, one of the many approaches for predicting binding pockets developed recently[11] can be applied for defining a pocket.

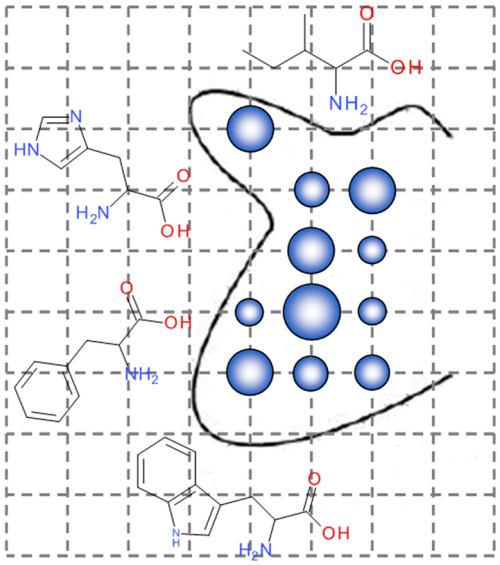


Figure 1. Computation of DrugScore potential fields within protein binding sites. To describe the preference of a protein binding site for a particular type of molecular interaction, a cubic grid is placed within the binding site. Then, a probe atom is placed at every grid point, and the strength of the interaction (denoted here by blue spheres of varying sizes) between this probe atom and all amino acids flanking the binding site is computed using the DrugScore scoring function.

Table 1. Probe Atoms and Their Associated Interaction Types

probe atom[a]	encoded interaction
C.ar	aromatic interactions
C.3	aliphatic interactions
O.2	hydrogen bond acceptor
N.3	hydrogen bond donor
O.3	hydrogen bond donor and acceptor

[a]The types of probe atoms follow the definition of types of ligand atoms used in DrugScore.[29]

Series Expansion in 3D Zernike Polynomials. Since the cubic, grid-based DrugScore potential fields are inappropriate for a fast and efficient computational analysis of unaligned protein binding sites, we decided to transform the fields to a simpler but still informative set of alignment-independent descriptors. To this end, we utilized a series expansion in 3D Zernike polynomials.[30] In general, this technique enables a compact description of any 3D object by a set of object-specific weights. It is based on the idea of generalized Fourier series, where a 3D object given as $f(\vec{x})$ is represented as a weighted sum of polynomials $Z_{n,l,m}(\vec{x})$ forming a complete orthogonal system of functions (eq 1):

$$f(\vec{x}) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega_{nl}^{m} \cdot Z_{nl}^{m}(\vec{x}) \quad (1)$$

To enable 3D function expansion in Cartesian coordinates, Canterakis[30] introduced the 3D Zernike polynomials (eq 2)

dx.doi.org/10.1021/ci300244y | J. Chem. Inf. Model. 2012, 52, 2339−2347

$$Z_{nl}^{m}(\vec{x}) = \left(\sum_{\vartheta=0}^{k} q_{kl}^{\vartheta} |\vec{x}|^{2\vartheta}\right) \cdot e_{l}^{m}(\vec{x}) \tag{2}$$

for $2k = n - l$. $e_l^m(\vec{x})$ are harmonic polynomials, and q_{kl}^{ϑ} are coefficients ensuring the orthonormality of the Zernike polynomials within the unit sphere. A detailed derivation and description of these Zernike polynomials can be found in the publication by Canterakis.[30]

The object-dependent coefficients Ω_{nl}^{m}, which are called Zernike moments, can be computed via eq 3:

$$\Omega_{nl}^{m} = \frac{3}{4\pi} \int_{|\vec{x}|\leq 1} f(\vec{x}) \cdot \overline{Z_{nl}^{m}(\vec{x})}\, \mathrm{d}\vec{x} \tag{3}$$

An algorithm for the fast numerical computation of these Zernike moments was presented by Novotni and Klein.[31] Finally, the Zernike moments can be transformed into rotation-invariant Zernike descriptors via eq 4:

$$d_{nl} = \sqrt{\sum_{m=-l}^{l} (\Omega_{nl}^{m})^2} \tag{4}$$

While any 3D object can be exactly described by a sum of infinitely many weighted Zernike polynomials according to eq 1, in practice, the series expansion is only performed for a limited number of Zernike polynomials (eq 5):

$$f(\vec{x}) \approx \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega_{nl}^{m} \cdot Z_{nl}^{m}(\vec{x}) \tag{5}$$

The parameter N is called the maximal expansion order and influences the amount of information about the 3D object that is encoded in the Zernike moments and the resulting alignment-independent Zernike descriptors. Equation 5 can also be used to compute reconstructed DrugScore potential fields based on the Zernike moments obtained from the original potential fields.

Computing Zernike Descriptors for Protein Binding Site Comparison. To compare protein binding sites based on DrugScore potential fields encoded by 3D Zernike descriptors, four major steps are required:

(A) Compute DrugScore potential fields for each binding site.
(B) Perform Zernike function expansion for each DrugScore potential field, i.e., compute the Zernike moments using eq 3.
(C) Transform the Zernike moments to the alignment-independent Zernike descriptors using eq 4.
(D) Compute the similarity of two binding sites in terms of the similarity of their corresponding Zernike descriptors (see below).

Data Sets. For validating the DrugScore-based Zernike descriptors, four data sets were used. Initially, five diverse proteins (Table 2) were used to investigate the information content of the Zernike descriptors as a function of the maximal expansion order N. Furthermore, one protein of this data set (PDB ID: 1b38) was used to show the alignment independence of the Zernike descriptors.

For evaluating the performance of the Zernike descriptors on a larger data set and comparing our approach to other techniques for binding pocket comparison, a benchmarking data set[22] for binding pocket comparison ("Hoffmann data set") was used. This data set contains 100 diverse proteins having a pairwise sequence identity <30%. The Hoffmann data set is composed of ten subsets, each of which contains ten proteins binding the same ligand. Thus, in this data set two proteins are assumed to be similar if they bind the same ligand. The ligands in the Hoffmann data set are of similar size, preventing a trivial discrimination of binding pockets solely based on the size of the binding site.

Table 2. Set of Five Diverse Proteins for Evaluating the Descriptive Power of the DrugScore-Based Zernike Descriptors

protein	abbreviation	PDB ID
cyclin-dependent kinase 2	CDK	1b38
carbonic anhydrase 2	CA	1oq5
HIV-1 protease	HIV	3o9a
adenosine A2a receptor	A2A	3qak
cyclooxygenase 2	COX	6cox

To investigate the performance of our Zernike descriptors in analyzing binding sites of proteins belonging to the same target family, we utilized a data set containing 23 ATP-binding sites of protein kinases ("protein kinase data set") from four different kinase subfamilies (cyclin-dependent kinase 2, glycogen synthase kinase 3 beta, lymphocyte-specific protein tyrosine kinase, p38 mitogen-activated protein kinase). In this data set two binding sites are considered similar if they belong to the same target subfamily. Thus, this data set provides a different view on binding site similarity compared to the Hoffmann data set. Furthermore, it allows an additional comparison of the Zernike descriptors to other approaches for binding site comparison because this data set was also used for the validation of FuzCav[14] and FLAP.[15]

To investigate the dependence of our Zernike descriptors on structural variations, we used a data set containing multiple crystal structures of 15 well-known drug targets assembled by Huang and Jacobsen[41] ("flexibility data set"). This data set contains proteins of varying flexibility, i.e., the largest movement of the DFG loop in three different p38 kinase structures is >10 Å, whereas the largest distance of corresponding side chains in two thrombin structures is <2 Å. Furthermore, this data set contains ligand-free crystal structures for four proteins.

Similarity Metrics. For evaluating our novel binding site descriptors, two different (dis)similarity measures are used. In order to compare a protein's original ($\vec{x}$) and reconstructed potential fields ($\vec{y}$), we used the Pearson correlation coefficients as a measure of reconstruction quality (eq 6):

$$r(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6}$$

This similarity measure was also used to compare Zernike moments and Zernike descriptors derived from different orientations of the same protein while analyzing the rotation (in)dependence of the descriptors. To this end, the Pearson correlation coefficient between Zernike moments or descriptors derived from MIFs of the original protein ($\vec{x}$) and Zernike moments or descriptors derived from MIFs of the rotated protein ($\vec{y}$) is computed.

To compare two pockets from different proteins in terms of the Zernike descriptors, we used the Manhattan distance (eq 7):

$$d(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i| \quad (7)$$

where $\vec{x}$ and $\vec{y}$ denote the Zernike descriptors derived for the two pockets to be compared.

Performance Criteria. To evaluate the ability of the DrugScore-based Zernike descriptors to detect similarities among related protein binding sites, we used two different performance criteria. For the Hoffmann data set, we used the area under the curve (AUC) as a measure of ranking performance because this measure was also used in the evaluation of other binding pocket comparison approaches on this data set.[22] To compute the AUC score for one query pocket, all other pockets in the data set are ranked according to their similarity to the query pocket in terms of the distance of the Zernike descriptors (eq 7). Then, a receiver operating characteristic (ROC) curve is computed showing the number of false positives (pockets binding another ligand than the query pocket) versus the number of true positives (pockets binding the same ligand as the query pocket). Finally, the area under this curve (AUC) is used as a performance criterion. AUC values vary between 0 and 1; an ideal ranking corresponds to AUC = 1, and a random ranking corresponds to AUC = 0.5.

When analyzing the protein kinase data set and the flexibility data set, we utilized Ward's hierarchical clustering implemented in R.[42] To perform the clustering, pairwise Manhattan distances of Zernike descriptors (eq 7) were used.

Availability. The source code to encode MIFs using Zernike descriptors is available upon request from H.G. for nonprofit research. DrugScore is available from the University of Marburg (http://pc1664.pharmazie.uni-marburg.de/download/ds).

RESULTS AND DISCUSSION

Reconstructing DrugScore Potential Fields. To initially evaluate the descriptive power of the DrugScore-based Zernike descriptors, we used five diverse proteins (Table 2) and performed series expansions for increasing maximal expansion orders. We selected an aromatic carbon atom as an exemplary probe atom and computed Zernike moments for the corresponding DrugScore potential fields using eq 3. Then, these Zernike moments were applied to compute reconstructed potential fields (eq 5). Finally, the reconstructed potential fields were compared to the original potential fields using Pearson's correlation coefficient as a measure of reconstruction quality. The higher the correlation between the original and reconstructed potential fields, the more information about the DrugScore potential fields is encoded in the Zernike moments and corresponding Zernike descriptors.

A maximal expansion order of 20 already yields a reconstruction quality of $r > 0.9$ (Figure 2). Furthermore, maximal expansion orders larger than 25 do not yield a significant improvement in reconstruction quality. Since similar results were obtained for other probe atoms too (data not shown), we decided to use $N = 25$ for the remainder of this study.

In Figure 3, original and reconstructed DrugScore potential fields for two exemplary proteins and an aromatic carbon probe atom are shown. The original potential fields consist of more than 50,000 grid points. By expanding the potential fields in terms of Zernike polynomials, this information is encoded by only 3276 Zernike moments from which the reconstructed potential fields were computed. Hence, the data is compressed to only 6.5% of the original size. The reconstructed potential fields also agree well with the original ones. This shows that a series expansion in Zernike polynomials is an elegant approach to transform 3D MIFs into computationally efficient and information-rich binding site descriptors.

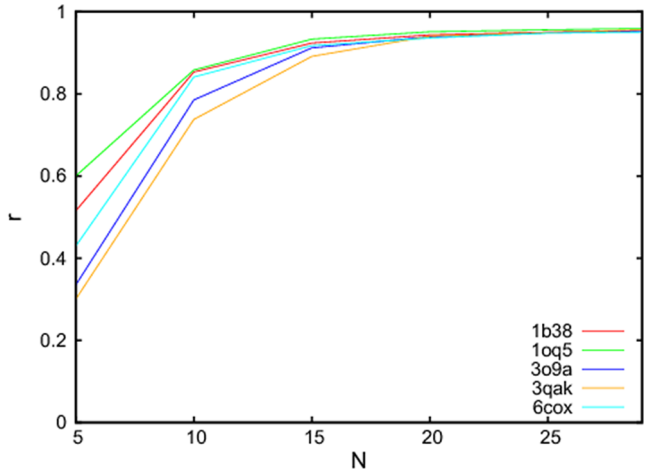


Figure 2. Reconstruction quality for different maximal expansion orders. For five diverse proteins, Zernike moments encoding DrugScore potential fields for the five probe atoms were computed and then used to reconstruct the potential fields. The correlation coefficient r (eq 6) between the original and the reconstructed potential fields, which is a measure of the reconstruction quality, is plotted for different maximum expansion orders N (eq 5). See Table 2 for abbreviations of the protein names.

Rotation Invariance of Zernike Descriptors. The Zernike moments computed via eq 3 can be transformed to a set of rotation-invariant and, thus, alignment-independent descriptors using eq 4. For $N = 25$, this leads to 182 Zernike descriptors per encoded potential field and, hence, a further data compression (data compressed to 0.4% of the original size). As a disadvantage of this transformation, the directional information of the 3D object is lost, thus making a reconstruction of the object solely based on the Zernike descriptors impossible. To show the alignment dependence of the Zernike moments and the alignment independence of the Zernike descriptors, we systematically rotated an exemplary protein (1b38) and then computed the resulting Zernike moments and Zernike descriptors for the different orientations. In Figure 4, the correlation coefficients (eq 6) between Zernike moments and Zernike descriptors, respectively, obtained for the rotated proteins and the initial orientation are shown. As expected, the Zernike moments vary significantly among the different orientations. By contrast, $r > 0.92$ is obtained for the Zernike descriptors, showing the alignment independence of these descriptors. The deviation from a perfect correlation can be explained by the fact that DrugScore potential fields are computed on a discrete, cubic grid within the binding pocket. Thus, different orientations of the same protein yield slightly different DrugScore potential fields; these differences are then carried forward to the Zernike descriptors.

Detecting Similarities among Diverse Proteins Binding Identical Ligands. To evaluate the performance of the DrugScore-based Zernike descriptors in detecting similarities among diverse proteins that bind identical ligands, we utilized the Hoffmann data set. Each protein was once used as a query, and AUC values were computed based on the ranking of all

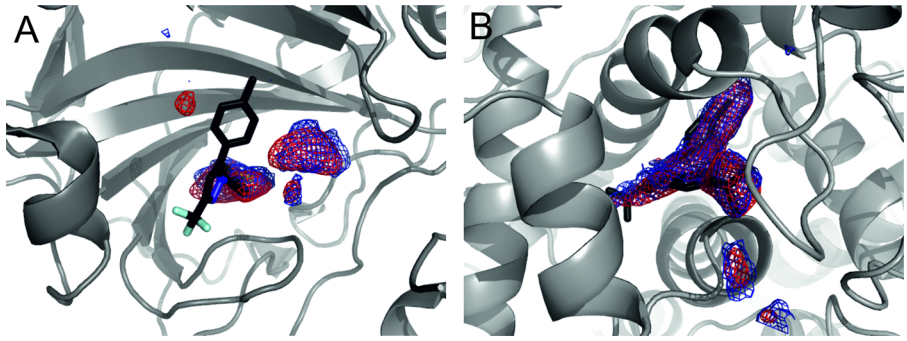


Figure 3. Comparing original and reconstructed DrugScore potential fields. DrugScore potential fields were computed for a probe atom of type aromatic carbon and two exemplary proteins (A, 1oq5; B, 6cox). Then, Zernike moments for a maximum expansion order of 25 were computed. The resulting Zernike moments were utilized to compute reconstructed potential fields. The original potential fields are shown in blue; the reconstructed ones are shown in red. All potential fields were contoured at a value of −20,000 DrugScore units.

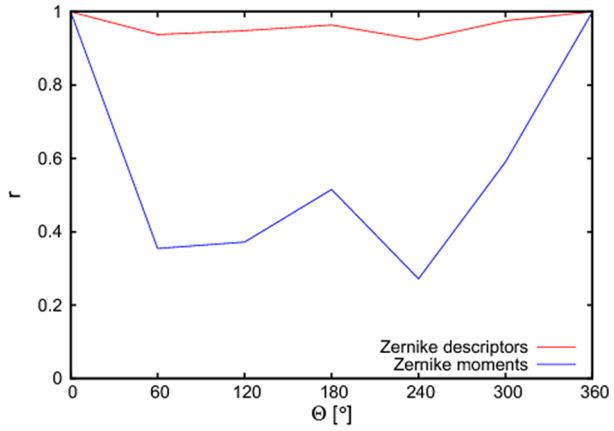


Figure 4. Verifying the rotation invariance of the Zernike descriptors. An exemplary protein (1b38) was systematically rotated around the *z*-axis, and Zernike moments and Zernike descriptors were computed for each orientation. To analyze the rotation dependence, correlation coefficients *r* (eq 6) between Zernike moments (red) and Zernike descriptors (blue), respectively, computed for each orientation and the initial one are plotted for different rotation angles Θ.

remaining proteins in the data set. The average AUC values for each of the ten ligand sets in the Hoffmann data set are shown in Table 3. The performance of our novel binding site descriptors varies among the different subsets: 7 out of 10 cases show an average AUC value >0.7 and 2 out of 10 cases <0.6. An almost perfect ranking is observed for proteins binding the ligand PMP, whereas the ranking is only slightly better than random for proteins binding the ligand 1PE. Interestingly, the performance of the Zernike descriptors encoding DrugScore potential fields for different probe atoms varies by <0.07 with respect to average AUC values within one subset of the Hoffmann data set. Consequently, the combination of Zernike descriptors encoding different DrugScore potential fields performs only marginally superiorly to the Zernike descriptors encoding only one individual probe atom. Thus, we conclude that the Zernike function expansion approach allows deriving information-rich binding site descriptors even for one individual probe atom, and that the performance of the resulting descriptors is almost independent of the selected probe atom.

Table 3. AUC Values for the Hoffmann Data Set[a]

ligand ID	C.ar	C.3	O.3	O.2	N.3	combined probe atoms
LDA	0.724	0.687	0.690	0.703	0.686	0.711
PMP	0.957	0.964	0.962	0.929	0.976	0.974
SAM	0.811	0.816	0.820	0.800	0.818	0.822
LLP	0.907	0.903	0.853	0.855	0.878	0.898
U5P	0.758	0.775	0.811	0.786	0.779	0.783
1PE	0.534	0.532	0.564	0.595	0.540	0.548
SUC	0.765	0.773	0.732	0.732	0.731	0.753
GSH	0.739	0.738	0.735	0.705	0.731	0.753
PLM	0.674	0.669	0.660	0.691	0.664	0.675
BOG	0.575	0.587	0.601	0.619	0.598	0.593
av	0.745	0.744	0.743	0.742	0.740	0.750

[a]Average AUC values for each set of proteins binding identical ligands are shown for Zernike descriptors encoding DrugScore potential fields for five individual probe atoms (see also Table 1), as well as the combination of all probe atoms.

In order to test to what extent the performance of the Zernike descriptors depends on the MIFs, we computed Zernike descriptors for EasyMIFs.[43] While DrugScore potential fields are computed using a knowledge-based scoring function, the EasyMIF potential fields are computed using the GROMOS force field and a distance-dependent dielectric.[44] Despite these differences, the performance of Zernike descriptors encoding DrugScore and EasyMIF potential fields is highly similar (Table 3 and Table S1 in the Supporting Information). As with DrugScore-based Zernike descriptors, the performance of Zernike descriptors encoding five different individual EasyMIF probe atoms, as well as the combination of all probe atoms, varies only slightly. Consequently, we conclude that the performance of Zernike descriptors does not critically depend on the MIF.

We decided to utilize the Hoffmann data set for validating the DrugScore-based Zernike descriptors because this data set was designed as a benchmarking data set for binding pocket comparison. In Table 4, the performance of the Zernike descriptors is compared to other methods for binding site comparison. Performance values of all other methods were taken from Hoffmann et al.[22] The two methods performing best on this data set are the sup-CK_L[22] approach (average AUC: 0.752) and our method (average AUC: 0.750), with the third best method (sup-CK) showing an average AUC value of 0.710. This result is remarkable because our approach does not require any parameter optimization or training. By contrast, the sup-CK_L approach has two parameters that were optimized particularly for the Hoffmann data set. Thus, our parameter-free

dx.doi.org/10.1021/ci300244y | J. Chem. Inf. Model. 2012, 52, 2339−2347

Table 4. Comparison of Different Methods for Binding Pocket Comparison[a]

method	AUC
sup-CK_L	0.752
Zernike descriptors	0.750
sup-CK	0.710
volume	0.648
MultiBind	0.690
sequence	0.577

[a]Reported are the performances of different methods for the Hoffmann data set in terms of AUC values averaged over all pockets. All AUC values expect for the Zernike descriptors were taken from Hoffmann et al.[22]

DrugScore-based Zernike descriptors perform as well as the optimal parameter set of the best performing method on this data set.

Furthermore, the results show that a similarity assessment based on protein sequences performs close to random (average AUC: 0.577). Thus, a structure-based analysis of protein binding sites using MIF-based descriptors uncovers similarities among proteins that cannot be detected based on sequence similarity.

Finally, we utilized the Hoffmann data set to investigate how the size of the grid box used for computing DrugScore potential fields influences the performance of the resulting Zernike descriptors. To this end, we computed DrugScore potential fields for grid boxes of varying size, i.e., extending by 3.5 Å, 3.75 Å, 4.0 Å, 4.25 Å, and 4.5 Å over the ligand in the crystal structure. Then, for each protein five individual trials were performed for which the size of the DrugScore grid box of the query protein was gradually varied according to the above values (3.5 Å to 4.5 Å) whereas the size of the DrugScore grid boxes of all other proteins was kept constant (4.0 Å). That way, the AUC values resulting from differently sized grid boxes of the query protein can be compared (Figure S1 in the Supporting Information). If the DrugScore grid box of the query protein and the ones of all other proteins have the same size (4.0 Å around the ligand), the average AUC value is 0.75. By incrementally increasing or decreasing the dimensions of the grid box of the query protein, only moderate decreases in the average ranking performances are observed (3.5 Å, AUC = 0.68; 3.75 Å, AUC = 0.72; 4.25 Å, AUC = 0.74; 4.5 Å, AUC = 0.71). Thus, we conclude that the performance of the Zernike descriptors is robust against moderate modifications of the grid box sizes. This fact is important if Zernike descriptors are used, e.g., for protein function prediction. Usually, novel proteins with unknown function are available only as ligand-free crystal structures; in that case, it will be more difficult to accurately place and define the grid box for computing DrugScore potential fields.

Clustering Protein Kinases. In the Hoffmann data set, two proteins are considered to be similar if they bind the same ligand. Another view on binding site similarity is given by the protein kinase data set consisting of 23 sites that bind ATP. In this data set, two binding sites are considered similar if they belong to the same target subfamily. Consequently, this data set can be used to assess whether an approach for binding site comparison is able to explain selectivities within a protein family. Again, Zernike descriptors for a maximal expansion order of 25 were computed for DrugScore potential fields of five different probe atoms. Then, Manhattan distances between the protein binding sites in terms of Zernike descriptors were used to cluster the protein kinase data set. The resulting cluster dendrogram obtained with five probe atoms (Table 1) is shown in Figure 5. The DrugScore-based Zernike descriptors lead to a perfect discrimination of the four different protein kinase subfamilies. This result also holds if any one of the individual probe atoms is used (data not shown).

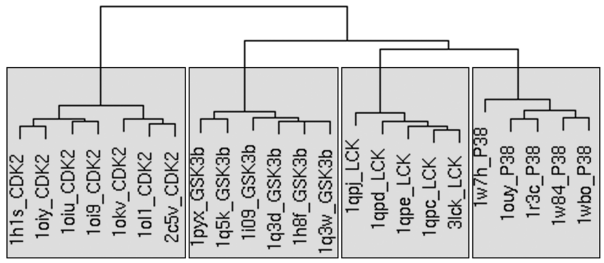


Figure 5. Clustering protein kinases. Hierarchical clustering was performed for 23 ATP-binding sites of four protein kinase subfamilies (GSK3b, glycogen synthase kinase 3 beta; CDK2, cyclin-dependent kinase 2; LCK, lymphocyte-specific protein tyrosine kinase, p38, p38 mitogen-activated protein kinase). Manhattan distances (eq 7) between Zernike descriptors encoding DrugScore potential fields for five probe atoms were used for similarity assessment.

This protein kinase data set was previously used to evaluate the performance of two other binding site comparison approaches: FuzCav[14] and FLAP.[15] Both approaches perfectly discriminated the different kinase subfamilies, too. Thus, again, our descriptors perform as well as other state-of-the-art techniques in the field.

Robustness of Zernike Descriptors with Respect to Structural Changes. Methods for analyzing protein binding sites are often hampered by conformational variabilities in the binding site.[45,46] To investigate to what extent the DrugScore-based Zernike descriptors tolerate such conformational variabilities, we used a set of multiple cocrystal structures of 15 drug targets[41] (Table S2 in the Supporting Information). As the data set also contains ligand-free crystal structures for four drug targets, it additionally allows investigating to what extent our DrugScore-based Zernike descriptors are able to detect similarities between *apo* and *holo* structures of the same protein. This is particularly important if it comes to predicting protein function for novel (usually ligand-free) protein structures, which is one major application field of binding site comparison.

We computed DrugScore-based Zernike descriptors for five probe atoms and then performed a hierarchical clustering based on Manhattan distances (eq 7). The resulting clustering dendrogram is shown in Figure 6. Notably, all crystal structures of one particular drug target form one cluster. These findings show that the Zernike descriptors are robust against conformational variabilities within the binding pocket. This result reflects that DrugScore potentials produce smooth binding energy landscapes[47,48] such that moderate conformational changes of the protein do not lead to gross changes in the potential values and, hence, neither to such changes in the Zernike descriptors. Finally, in all cases, *apo* and *holo* structures of one drug target were correctly grouped together (Figure 6), showing that our Zernike descriptors were successful also in the case of ligand-free binding sites.

Computed Manhattan distances of Zernike descriptors are given in Table S2 in the Supporting Information for each pair of crystal structures of one protein. These distances almost

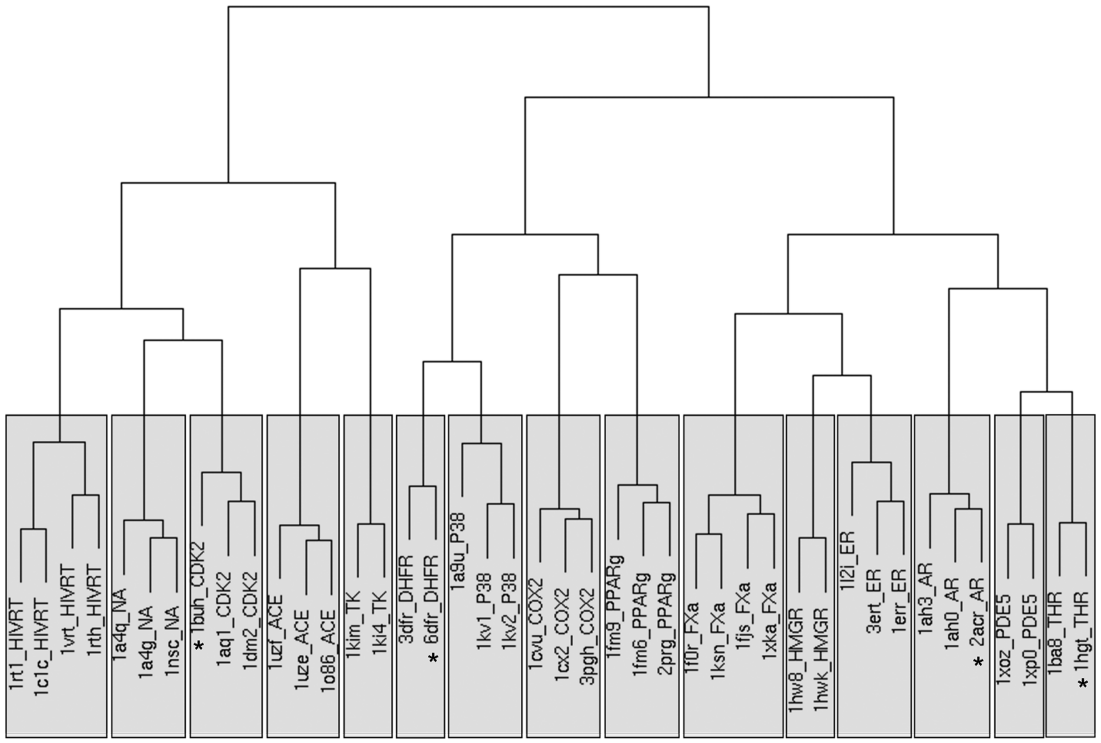


Figure 6. Clustering the flexibility data set. Hierarchical clustering was performed for multiple crystal structures of 15 drug targets showing different degrees of side chain movements in the binding site. See Table S2 in the Supporting Information for abbreviations of the names of the drug targets. Manhattan distances (eq 7) between Zernike descriptors encoding DrugScore potential fields for five probe atoms were used for similarity assessment. *Apo* structures are marked by an asterisk.

always reflect the magnitude of the conformational variabilities between the binding sites in terms of maximal rmsd between two corresponding side chains or loops (see also Figure S2 in the Supporting Information), resulting in a significant correlation between the magnitude of conformational variabilities and the difference in terms of Zernike descriptors (R^2 = 0.523). The largest distances between Zernike descriptors for different crystal structures of one drug target are observed for the three p38 protein kinases, which also exhibit the largest loop movement (up to 11.26 Å) within the binding site. We conclude that Manhattan distances of Zernike descriptors reflect the extent of conformational variability observed in the binding site of a protein.

CONCLUSIONS AND OUTLOOK

In this study, we have introduced novel alignment-independent descriptors for protein binding site analysis. We characterize protein binding sites by DrugScore potential fields, which are then transformed to a set of information-rich descriptors using a series expansion in 3D Zernike polynomials. The resulting DrugScore-based Zernike descriptors showed a promising performance in detecting similarities among diverse proteins binding identical ligands. Thus, the descriptors are a suitable tool for analyzing binding sites that are likely to bind identical or similar ligands. This is important when it comes to predicting potential off-targets for novel drug candidates or novel targets for known drugs.

Furthermore, the DrugScore-based Zernike descriptors allowed a perfect clustering of ATP binding sites from different protein kinase subfamilies. Thus, the Zernike descriptors allow analyzing target families based on the similarity of the binding sites, which is important in the development of drugs with high selectivity.

Finally, the DrugScore-based Zernike descriptors are robust against structural variations within the binding sites. In addition, a comparison of *apo* and *holo* structures of identical proteins demonstrated that the Zernike descriptors can be applied to characterize ligand-free binding pockets, too. This is of crucial importance for applying the Zernike descriptors to structure-based protein function prediction because structural information of proteins with unknown function is usually only available without ligand information.

In summary, encoding DrugScore potential fields by Zernike descriptors enables a fast and efficient similarity assessment between pairs of protein binding sites. In addition, we expect the Zernike descriptors to be useful also in other areas of binding site analysis, e.g., for druggability predictions.

ASSOCIATED CONTENT

Supporting Information

Tables with AUC values for the Hoffmann data set using EasyMIF potential fields (Table S1) and information for the flexibility data set (Table S2) as well as figures showing the influence of varying DrugScore grid box sizes on Zernike descriptor performance using the Hoffmann data set (Figure S1) and structural differences between binding sites versus Zernike distances using the flexibility data set (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-211-81-13662. Fax: +49-221-81-13847. E-mail: gohlke@uni-duesseldorf.de.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5*, 2438−2452.

(2) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

(3) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2009**, *23*, 209−219.

(4) Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123−134.

(5) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrases by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *4*, 550−557.

(6) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistance tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e10000423.

(7) De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One* **2010**, *5*, e12214.

(8) Miletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418−1431.

(9) Tseng, Y. Y.; Dundas, J.; Liang, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.* **2009**, *387*, 451−464.

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

(11) Leis, S.; Schneider., S.; Zacharias, M. In silico prediction of binding sites on proteins. *Curr. Med. Chem.* **2010**, *17*, 1550−1562.

(12) Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K. Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011**, *51*, 2666−2679.

(13) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.

(14) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(15) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and applications. *J. Chem. Inf. Model.* **2007**, *47*, 279−294.

(16) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(17) Chikhi, R.; Sael, L.; Kihara, D. Real-time ligand binding pocket database search using local surface descriptors. *Proteins* **2010**, *78*, 2007−2028.

(18) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 72−730.

(19) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **2007**, *386*, 283−301.

(20) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.

(21) Xie, L.; Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence-order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5441−5446.

(22) Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *11*, 99.

(23) Cross, S.; Cruciani, G. Molecular fields in drug discovery: getting old or reaching maturity. *Drug Discovery Today* **2010**, *15*, 23−32.

(24) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(25) Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033−3044.

(26) Naumann, T.; Matter, H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J. Med. Chem.* **2002**, *45*, 2366−2378.

(27) Pirard, B. Peroxisome proliferator-activated receptors target family landscapes: a chemometrical approach to ligand selectivity based on protein binding site analysis. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 785−796.

(28) Angel, D.; Martinez, G. C.; Pastor, M. Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in molecular interaction fields. *J. Chem. Inf. Model.* **2008**, *48*, 1813−1823.

(29) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(30) Canterakis, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. *Proceedings of the 11th Scandinavian conference on image analysis*, Kangerlussuaq, Greenland, 1999; pp 85−93.

(31) Novotni, M.; Klein, R. Shape retrieval using 3D Zernike descriptors. *Comput.-Aided Des.* **2004**, *36*, 1047−1062.

(32) Grandison, S.; Roberts, C.; Morris, R. J. The application of 3D Zernike moments for the description of "model-free" molecular structure, functional motion, and structural reliability. *J. Comput. Biol.* **2009**, *16*, 487−500.

(33) Venkatraman, V.; Sael, L.; Kihara, D. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell. Biochem. Biophys.* **2009**, *54*, 23−32.

(34) Mak, L.; Grandison, S.; Morris, R. J. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *J. Mol. Graphics Model.* **2008**, *26*, 1035−1045.

(35) Venkatraman, V.; Chakravarthy, P. R.; Kihara, D. Application of 3D Zernike descriptors to shape-based ligand similarity searching. *J. Cheminf.* **2009**, *1*, 1−19.

(36) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(37) Sotriffer, C. A.; Gohlke, H.; Klebe, G. Docking into knowledge-based potential fields: a comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45*, 1967−1970.

(38) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(39) Pfeffer, P.; Gohlke, H. DrugScore^{RNA}—knowledge-based scoring function to predict RNA-ligand interactions. *J. Chem. Inf. Model.* **2007**, *47*, 1868−1876.
(40) Krüger, D. M.; Gohlke, H. DrugScore^{PPI} webserver: Fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* **2010**, *38*, W480−W486.
(41) Huang, N.; Jacobson, M. P. Binding-site assessment by virtual fragment screening. *PLoS One* **2010**, *5*, e10109.
(42) R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. Available at: http:www.R-project.org.
(43) Ghersi, D.; Sanchez, R. EasyMIF and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25*, 3185−3186.
(44) Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(45) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455−476.
(46) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sortriffer, C. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237−6255.
(47) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *44*, 2287−2303.
(48) Kazemi, S.; Krüger, D. M.; Sirockin, F.; Gohlke, H. Elastic potential grids: accurate and efficient representation of intermolecular interactions for fully-flexible docking. *ChemMedChem* **2009**, *49*, 181−188.