JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Combining Global and Local Measures for Structure-Based Druggability Predictions

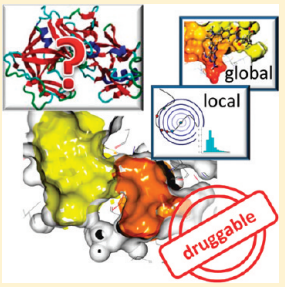
Andrea Volkamer,[†] Daniel Kuhn,[‡] Thomas Grombacher,[§] Friedrich Rippmann,[‡] and Matthias Rarey*,[†]

[†]University of Hamburg, Center for Bioinformatics, Bundesstr. 43, 20146 Hamburg, Germany

[‡]Merck KGaA, Merck Serono, Global Computational Chemistry, Frankfurter Str. 250, 64293 Darmstadt, Germany

[§]Merck KGaA, Merck Serono, Bioinformatics, Frankfurter Str. 250, 64293 Darmstadt, Germany

ABSTRACT: Predicting druggability and prioritizing certain disease modifying targets for the drug development process is of high practical relevance in pharmaceutical research. DoGSiteScorer is a fully automatic algorithm for pocket and druggability prediction. Besides consideration of global properties of the pocket, also local similarities shared between pockets are reflected. Druggability scores are predicted by means of a support vector machine (SVM), trained, and tested on the druggability data set (DD) and its nonredundant version (NRDD). The DD consists of 1069 targets with assigned druggable, difficult, and undruggable classes. In 90% of the NRDD, the SVM model based on global descriptors correctly classifies a target as either druggable or undruggable. Nevertheless, global properties suffer from binding site changes due to ligand binding and from the pocket boundary definition. Therefore, local pocket properties are additionally investigated in terms of a nearest neighbor search. Local similarities are described by distance dependent histograms between atom pairs. In 88% of the DD pocket set, the nearest neighbor and the structure itself conform with their druggability type. A discriminant feature between druggable and undruggable pockets is having less short-range hydrophilic−hydrophilic pairs and more short-range lipophilic−lipophilic pairs. Our findings for global pocket descriptors coincide with previously published methods affirming that size, shape, and hydrophobicity are important global pocket descriptors for automatic druggability prediction. Nevertheless, the variety of pocket shapes and their flexibility upon ligand binding limit the automatic projection of druggable features onto descriptors. Incorporating local pocket properties is another step toward a reliable descriptor-based druggability prediction.



INTRODUCTION

The ultimate goal in pharmaceutical research is to find novel drugs against a specific disease. The drug discovery process is very time-consuming and costly. A survey from 2004 estimated the average time for drug development to be more than 12 years and costs around $1150 billion U.S.[1] Although the total number of human genes is high, only 10% of the human genome is involved in disease onset and progression, enabling a set of 3000 potential targets.[2] Interestingly, data from 2003 showed that 60% of drug discovery projects fail because the underlying target was found to be not druggable.[3] Therefore, it is of the utmost importance to prioritize those targets that have a high chance of being amenable to the drug development process.

Target assessment involves multiple parameters considering disease relevance, structural aspects, screening feasibility, selectivity, and toxicology considerations, as well as commercial attractivity. One prerequisite of a druggable target is its general ability to be addressed by low-molecular weight compounds. Dealing with this challenge, the term druggability has been coined.[2,4−6] Since the term druggability covers many different optimization parameters from early to late stage drug discovery, several definitions have been proposed. In 2002, the term druggability was defined as the potential of a disease-modifying target to be modulated by orally bioavailable drug compounds.[2] To describe the general potential of a known binding pocket to interfere with small molecules, the terms targetability, ligandability, or chemical tractability [1,7−9] have likewise been introduced. Besides the druggability prediction of previously known binding pockets, the identification of putative allosteric sites is of high practical value. Allosteric sites may provide an additional way to modulate protein function, especially in cases where the active sites of related proteins are almost identical. The challenge is to rapidly and reliably estimate the chances that the activity of a target can be sufficiently modulated by a small molecule.

Over the past 15 years, druggability predictions have been a field of active research, and a variety of methods have been developed. The early methods relied on experimental druggability assessment,[5,10,11] e.g., NMR-based fragment screening performed by Hajduk and co-workers.[10] NMR-based fragment screening for druggability predictions are still broadly used in pharmaceutical research, and hit rates detected in such NMR screens correlate with a chance of success in hit-to-lead programs.[8] Protein-observed NMR screening is able to detect very weak binders and additionally provides information about the actual binding site of the ligands.[10] It requires, however, a soluble and medium-sized protein (less than 50 kDa) that could be expressed in *E. coli* for the isotope labeling. Ligand-observed NMR screening (e.g., WaterLOGSY (water ligand optimized

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: September 23, 2011
Published: December 8, 2011

Table 1. Automatic Methods for Pocket and Druggability Prediction

method	pocket detection	druggability score	# proteins
SiteMap[17] 2009	VdW energies and buriedness on a 3D grid	DScore = $0.094(n)^{1/2} + 0.6e - 0.324p$ n = number of site points, e = enclosure, p = hydrophilicity	17 druggable, 6 difficult, and 4 undruggable
fpocket[18,21] 2009/2010	α spheres	drugscore = $e^{-z/(1+e-z)}$ $z = \beta_0 + \beta_1 f_1(d_1) + \beta_2 f_2(d_2) + \beta_3 f_3(d_3)$ d_1 = local hydrophobicity density d_2 = hydrophobicity score d_3 = normalized polarity score	NRDD: 5 prodrug and 20 undruggable DD: 919 druggable, 67 prodrug, and 84 undruggable
DoGSiteScorer[20] 2010/2011	Difference of Gaussian on a 3D grid	support vector machine (libsvm), similarity search (nearest neighbor)	NRDD, DD

gradient spectroscopy) or STD (saturation transfer difference)) works for larger proteins and requires no isotope labeling.[12] However, the application of NMR screening may be hampered by the lack of a sufficient amount of protein available, since druggability screens are typically performed very early in the project. Hence, computational models have been constructed that require less resources. Due to the large amount of available genome data, the first approaches have been dominated by sequence-based methods. Machine learning techniques working on descriptors solely derived from sequence information were used to predict the druggability of targets.[1,13−15] According to a review about target druggability from 2008,[1] prediction accuracies of sequence-based methods lie around 68%,[13,15] suggesting the need of additional descriptors or measures. Due to the rising number of solved target structures, the development of structure-based computational methods was recently stimulated (e.g., Map_{POD},[11] SCREEN,[16] DLID,[7] SiteMap,[17] fpocket[18]). A fundamental step in high-throughput druggability assessment is a reliable and accurate method for detecting the ligand binding sites.[1,19] Whereas many methods exists that are able to detect ligand binding sites,[19,20] methods that simultaneously estimate the druggability of a protein solely from its structure in an automated manner have been developed only recently. SiteMap,[17] fpocket,[18,21] and the newly introduced method DoGSiteScorer fall into the latter category (Table 1). Due to the automatic pocket detection step, the complexity of this task is increased. If a detection algorithm misinterprets a cavity as a binding pocket or wrongly assigns the pocket boundary, clearly the druggability prediction must fail as well. All three algorithms predict pockets solely based on the atomic coordinates of the protein. SiteMap uses van der Waals (VdW) energies and a buriedness value calculated on a grid to predict pockets on the protein surface. Fpocket investigates α spheres for active site predictions. In DoGSiteScorer, pockets and subpockets are predicted with DoGSite[20] using a Difference of Gaussian filter.

Different types of in silico methods have been employed to select and characterize the most promising target candidates. Molecular docking, e.g., is used as predictive tool to estimate the druggability of a given protein,[1,9] using success rates of docking drug-like ligands into its pocket. Aside from docking-based approaches, most existing computational methods use descriptors derived in the active site in combination with techniques from statistical learning to distinguish between druggable and undruggable targets. Statistical learning approaches have three requirements: a description of the discriminating properties, a classification algorithm, and a labeled data set for training purpose.

No matter how the method works, it is vital for reliable predictions to have a large and unbiased data set on which to evaluate the method. This point is a well-known pitfall in the druggability prediction field. Collecting positive examples is rather simple, since only targets with known orally available drugs have to be selected. Concerning the negative data set, such a statement is more difficult. While the terms druggability, ligandability, and targetability agree that the target has to be able to accommodate a ligand, they disagree in the features of the ligand, ranging from being small, to being drug-like and highly potent, to additionally being orally bioavailable. Following the definition of Cheng et al.,[11] targets that do not bind an orally bioavailable drug are classified as being undruggable. This includes pockets that do not bind any small molecule and pockets that bind a small molecule which is not orally bioavailable or not drug-like. But the question remains whether either orally bioavailable drugs do not exist for the respective target or they have not been identified so far. Using NMR-based fragment screens, Hajduk and co-workers published in 2005 a first druggability data set consisting of 23 structures.[10] A few years later, Cheng et al. published a subsequent set with 63 structures of 27 pharmaceutical targets evaluated on a desolvation free energy model.[11] Very recently, Schmidtke and Barril[21] published a freely available druggability data set (DD) and a nonredundant version of it (NRDD). For the positive data set, proteins in complex with an orally bioavailable marketed drug have been collected from Vieth et al.[22] and verified with the aid of DrugBank.[23] After several filter steps and manual inspection, this data is introduced into the DD classified as druggable, difficult (in the case of prodrugs), or undruggable, e.g., if the compound lacks in bioavailability. Furthermore, the data sets of Cheng et al.[11] and Hajduk et al.[10] were added. In total, the DD consists of 919 druggable, 67 prodrugs, and 84 undruggable targets, showing the evident bias toward druggable data. To eliminate redundancies and bias toward specific protein families in the data, a 70% sequence cutoff was used to create the NRDD data set, consisting of 45 druggable, 5 prodrug, and 20 undruggable proteins.

To apply a classification algorithm, a description of pocket properties discriminative with respect to druggability is needed. Thus, relevant structural, geometrical, and physicochemical features are identified from known protein−ligand complexes.[10,11,16,17,21] These descriptors are then used to predict the druggability of newly identified targets. Summarizing these studies, there is agreement about the properties that are important in distinguishing between druggable and undruggable pockets: namely, size, shape complexity, and hydrophobicity. Hajduk et al.[10] correlated 13 binding site characteristics with hit rates of 23 targets. A regression analysis showed that a conjunction of eight properties was most discriminative. While no single property was able to distinguish between the classes, they stated that most important for the druggable

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

character was a higher apolar surface area. In SCREEN, developed by Nayal and Honig,[16] 99 proteins were studied, and a feature vector of 408 descriptors was established. A decision-tree-based algorithm was incorporated, finding a combination of 18 descriptors having the highest predictivity. Pocket size and shape were considered as being more important than physicochemical properties. In the work of Cheng et al.,[11] pocket curvature and the fraction of nonpolar solvent accessible surface area are combined with properties of drug-like ligands. Map_{POD}, a model to assess the maximum affinity of a hypothetical drug-like ligand, was developed. In this desolvation-based free energy model, the druggability is mostly described by the hydrophobic effect. In 2009, Weisel et al. extended PocketPicker,[24] an automated shape-based pocket-prediction method, to estimate the ability of a pocket to accommodate small drug-like molecules. Druggability is predicted by autocorrelation vector-based shape descriptors of potential binding sites. Self-organizing maps are used for the classification of 210-dimensional shape descriptors, comprising size and buriedness of a pocket. SiteMap,[17] published in 2009, predicts a site score (SiteScore) and a druggability score (DScore) through a linear combination of only three single descriptors: pocket volume encoded by site points, hydrophobicity, and shape (Table 1). The two scores differ in the coefficients, which are based on different training sets and strategies. The coefficients of SiteScore were calculated using regression analysis on the PDBbind proteins[25] and the DScore on a test set of 69 targets.[11] In 2010, Sheridan et al. published DLID,[7] a method to estimate the bindability of pockets for the whole PDB based on the drug-like density. DLID measures the likelihood of a pocket to bind a drug-like molecule. Therefore, the number of pockets containing drug-like ligands in the local neighborhood in pocket space is calculated. DLID is predicted using a linear regression on only three pocket properties: logarithmic volume, buriedness, and hydrophobicity of the pocket. Schmidtke and Barril[21] selected for their DrugScore mainly physicochemical features normalized by size as druggability predictors which are combined into an exponential function. A bootstrapping method is used to choose those parameters yielding the highest accuracy on the NRDD (Table 1). The mean local hydrophobic density of the binding site is considered as the most predictive descriptor combining the size and spatial distribution of hydrophobic agglomerations into a single number.

While in the past most approaches focused on global pocket properties, recently a shift toward local descriptions has been observable. Although a druggable pocket usually has a slightly hydrophobic character, polar interactions present necessary anchor positions. They play a fundamental role in binding soluble compounds as well as directing a ligand with high specificity.[1] Schmidtke and Barril investigated the effect of the local environment on the energetics of association of polar groups. The change in accessible surface area (ASA) is analyzed as a function of the radii used to represent the atoms. For druggable pockets, the change in polar surface area is significantly smaller than the change in apolar surface area. Nevertheless, this atomic detail information is not used as a parameter in DrugScore. The change in polar and apolar ASA was found to be highly dependent on variability in active site predictions and not applicable to high-throughput predictions. The ASA change provides a finer level of detail and may be better suited for individual applications such as predicting the druggability of known binding sites.

In this work, a new fully automated pocket and druggability prediction approach called DoGSiteScorer (Table 1) is introduced. The recently published pocket detection algorithm DoGSite[20] is used to calculate over 40 global descriptors per target pocket. With the aid of a shrinkage discriminant analysis, a small set of descriptors is selected encoding important properties of druggable pockets. These descriptors are used in DoGSiteScorer to train a support vector machine (SVM) model for druggability predictions, based on druggable and undruggable pockets from the NRDD. For direct comparison to previously published methods,[7,17] a second SVM model including decoy pockets is introduced. Additionally, a simple linear score (SimpleScore) is calculated on the basis of a regression analysis on three descriptors encoding size, shape, and hydrophobicity. The largest available druggability data set (DD[21]) is used to evaluate DoGSiteScorer. Furthermore, local properties are investigated to better describe the features necessary for ligand binding, without being dependent on the overall shape of the pocket or its boundary. To emphasize local pocket features, functional groups present in the pockets are analyzed in terms of distance-dependent histograms between atom pairs. Finally, these profiles are used in a homology-based nearest-neighbor search taking pocket similarities into account. Such a neighborhood analysis can especially bring the annotation of flexible or intermediate targets forward which only show partial similarity to known drug targets. Analyzing the local pocket properties is another step toward a reliable descriptor-based druggability prediction.

MATERIAL AND METHODS

Data Preparation. In this work, the druggability data set (DD) as well as the nonredundant version of it (NRDD) compiled by Schmidtke and Barril[21] are used. The DD contains in total 1069 structures, which are divided into 10 classes with increasing druggability character (Figure 1). To avoid artificial

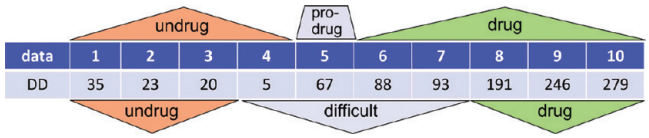


Figure 1. Statistics of DD structures, separated into 10 druggability levels. The labels above the table specify the fpocket division; the labels below the table, the grouping used in this work.

classification results that could arise from a too granular input categorization, the number of classes is reduced. In the original DD paper, a classification into the three classes undruggable, prodrug, and druggable is provided. Only the 67 prodrug structures are originally assigned to the intermediate class. Nevertheless, uncertainties remain in the classification of borderline druggable (6,7) and undruggable (4) structures, which may cause misclassifications. After personal communication with the authors, the DD data set is in this work separated into the categories druggable, difficult, and undruggable, as shown in Figure 1, providing a more balanced distribution.

DoGSiteScorer is directly compared to the previously published methods fpocket[21] and SiteMap.[17] To allow for a fair comparison, the NRDD is prepared as described by Schmidtke and Barril. DoGSite provides two levels of granularity. The description given here refers to the pocket level, while the procedure is similarly performed on the subpocket level. First, all pockets are predicted for the 70 NRDD proteins. The respective pocket is picked, on the

basis of the contained drug or nondrug, specified by Schmidtke and Barril. If unspecific or wrong pockets are predicted, noise is added to the subsequent druggability prediction. Therefore, similar to the mutual overlap criterion used in fpocket and a ligand atom distance criterion applied for SiteMap, we incorporate the DoGSite coverage criterion[20] to restrict the considered pockets. Only pockets with over 40% ligand and 20% pocket coverage are regarded for druggability predictions. These pockets constitute the NRDD data set. To enrich the negative data set, so-called decoys are additionally introduced into a second data set (decoyNRDD). Decoys are predicted pockets with a volume of at least 100 Å^3 not containing any ligands. Note that due to the individual pocket prediction step of each method, neither the pockets nor the decoys used in fpocket, SiteMap, or DoGSiteScorer are identical. Table 2 comprises statistics about druggable, undruggable, and decoy pockets for each method.

Table 2. Overview of the Number of Pockets Contained in the NRDD and the decoyNRDD Data Set

method	drug	undrug	decoy	NRDD	decoyNRDD
DoGSite(Poc)	67	22	328	89	417
DoGSite(SPoc)	72	24	331	97	428
fpocket	70	16	354	89	440
SiteMap	63	367		–	430

Besides enrichment studies on the decoyNRDD, the performance of the method on the complete DD is investigated. This data set comprises holo and apo structures. In order to be able to compare the potential ligand binding pockets of apo structures, they are superposed onto their respective holo structure, and the ligand is copied. The mapping of apo and holo structures is performed on the basis of the target classification in the DD. For the analysis, pockets for all DD structures are predicted and all ligand containing pockets fulfilling the coverage criterion are investigated. In total, our DD data set comprises 901 entries: 649 druggable, 208 difficult, and 44 undruggable pockets, respectively.

Method for Pocket Prediction and Descriptor Calculation. Protein pockets are predicted with DoGSite. The protein is mapped onto a grid and a Difference of Gaussian filter is used to identify pockets on the protein surface. The predicted binding pockets can additionally be split into subpockets which describe the ligand accessible volume more accurately. A detailed description of the pocket prediction algorithm can be found in the previous DoGSite publication.[20]

For each predicted binding pocket, several global and local descriptors are calculated. The pocket-forming grid points and the pocket-lining residues are the basis for the calculation of the different pocket properties. Spatial properties of a pocket are represented by pocket volume, surface, and depth. Hence, all pocket grid points have to be labeled according to their environment (Figure 2). Per default, each point in the pocket is considered as a volume grid point. In 3D grid space, each point is by default surrounded by 26 grid points. Since the pocket volume is continuous, each inner grid point has exactly 26 pocket dedicated neighbors. Grid points at the pocket border have less than 25 neighbors and are defined as hull grid points. All hull grid points that are within a specific distance of any protein atom are labeled as surface, the remaining as solvent exposed grid points. The atom distance cutoff is set to *atom_vdW_radius* + 1.2 Å (the radius of a hydrogen atom),

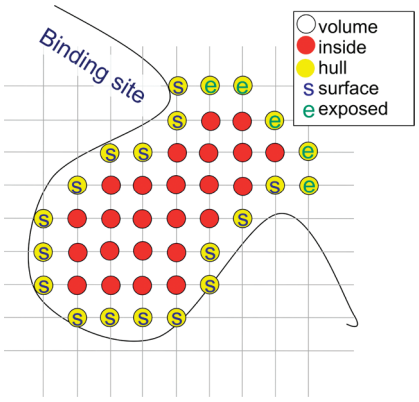


Figure 2. Two dimensional schematic view of grid point assignment.

such that the solvent-exposed grid points relate to the part of the pocket that is accessible to water.

Volume is discretely described as the number of volume grid points multiplied by the cubed grid spacing (Figure 3). Surface is calculated as the number of surface grid points multiplied by the squared grid spacing. Binding pockets are further characterized by their buriedness using a breadth-first search (BFS) starting from the solvent-exposed points of the pocket (Figure 3). The depth is described by the maximal separation between solvent and buried part of the pocket. Therefore, the maximal distance between two points, one from the solvent exposed shell and one from the deepest BFS shell, is calculated. If a pocket is completely buried, the farthest distance between two hull grid points is calculated, describing the maximal diameter of the pocket.

Besides volume, shape is another important characteristic for druggable sites. To capture the shape of a pocket, ellipsoids are fitted into the pocket (Figure 3). The ellipsoidal main axes a, b, and c are calculated and sorted by length: $a > b > c$. The relation between the lengths of those axes describes the shape being something between a sphere ($a \approx b \approx c$), a disk ($a \approx b > c$), and a rod ($a > b = c$). In addition, the complexity or the roughness of the pocket is described. The ratio of surface to volume grid points (gps_s_v) and the portion of the lid of the pocket compared to the hull (gps_se_h) are analyzed. The gps_se_h quotient describes the enclosure; it is higher if the pocket is rather open. The proportion of fitted ellipsoid volume to discrete pocket volume (ellips_vol) is another indicator for pocket complexity.

Physico-chemical properties of the binding site are covered by analyzing pocket atoms in terms of type, functional groups, and amino acids to which they belong. Protein atoms that lie within *atom_vdW_radius* + 1.2 Å of any pocket grid point are considered as pocket atoms. The number of pocket atoms per element type (carbon, oxygen, nitrogen, sulfur, and others) are calculated and normalized by the total number of pocket atoms. In addition, the number of hydrophilic anchor points, in terms of donor and acceptor atoms present in the pocket, is assessed. In the same way, the amino acids to which these atoms belong are investigated and analyzed by type. Furthermore, they are grouped into positive, negative, polar, and apolar amino acid subsets. Finally, metals found in the pocket are counted. To describe the overall lipophilic character of the pockets, so-called site interaction centers (SIACs) [26,27] are calculated for each atom. SIACs specify the interaction profile of the pocket atoms with a potential ligand and have been successfully applied in docking studies.[28] As introduced by the FlexX[28] model, functional groups are described by interaction centers and

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

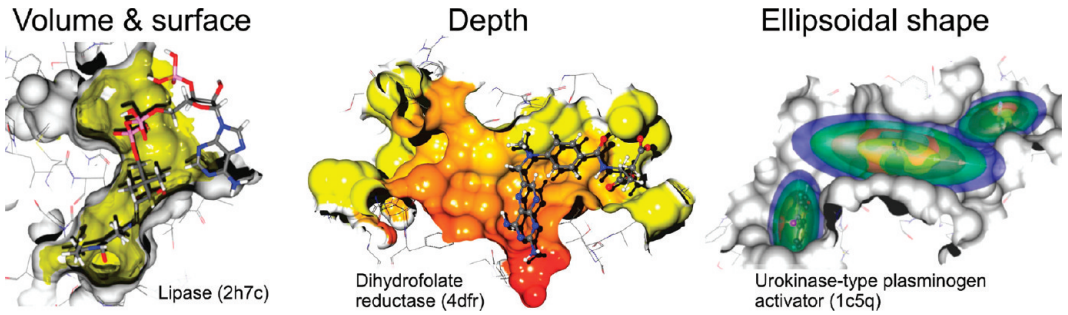


Figure 3. Example of global descriptors based on pocket grid points, from left to right: volume and surface, depth, and fitted ellipsoids. The protein surface is drawn in gray; the volume of the pocket is drawn in yellow. The depth of the pocket is projected onto the surface, color coded from yellow to red.

surfaces. These surfaces are continuously discretized by a number of evenly distributed interaction dots. SIACs are a coarse-grained representation of these dots, grouped into three types: hydrogen bonds, metal coordinations, and lipophilic contacts. The number of lipophilic SIACs in the pocket divided by the number of lipophilic and hydrophilic SIACs (including hydrophilic metal interactions) constitute the lipophilic character (lipo_siac). Furthermore, the lipophilic surface character is described by calculating the lipophilic solvent accessible surface (SAS) divided by the complete SAS of the pocket (lipo_surf).

Besides these global descriptors, local features are analyzed by the introduction of a local functional group profile. In this profile, distances between pairs of functional group atoms, namely hydrophilic−hydrophilic and lipophilic−lipophilic atom pairs, are calculated. A histogram separated in 2 Å bins is compiled for each single functional atom in the pocket. Starting from each atom, a radial search is performed. In Figure 4, this is

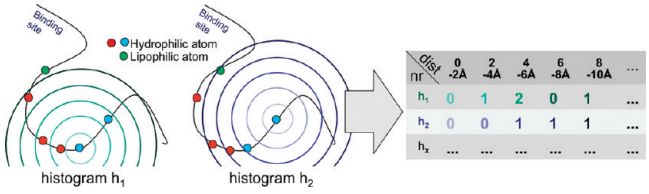


Figure 4. Schematic view of distance-dependent pair histogram calculation.

exemplarily shown for two functional atoms. In the case of h_1, no partner is found within 0−2 Å; one is found between 2−4 Å, another two between 4−6 Å, and so on. This procedure is repeated for every single functional atom in the pocket. Eventually, each line of the histogram table describes the local chemical neighborhood of a single atom in the pocket.

Druggability Prediction. To predict the druggability of binding pockets, two strategies are pursued: a machine learning approach based on global descriptors and a nearest neighbor approach based on local features.

A. Global Machine Learning Approach.

Machine Learner Selection. There are several machine learning techniques—such as Bayesian nets (BN), random forests (RF), and support vector machines (SVM), that have been successfully used to classify complex data sets. The statistical analysis software Orange[29] is employed to test different machine learning methods against each other to detect the method which performs best and to overcome any potential short-coming of one classifier. Predictions were made using the Orange prediction learner with default settings for BN, RF, and SVM on the decoyNRDD, as shown on the left side of Figure 5. A 25-fold random sampling on 50% relative training test size is performed. Receiver operating characteristic (ROC) curves and areas under the curves (AUCs) are provided. All three methods have classification accuracies above 81%. The average ROC curves on the right side of Figure 5 show that SVM slightly outperformed BN and RF, with AUCs of 70%, 59%, and 57%, respectively, and became our method of choice.

Note that Orange was only deployed to choose the learning method. In DoGSiteScorer, the freely available SVM software package libsvm[30] is used. This SVM requires as input a list of pockets with the aforementioned calculated descriptors and a predefined druggability class for the pockets in the training set. Furthermore, the calculation of a confidence value is enabled. Besides the classification of any query target into druggable and undruggable, a druggability score between zero (undruggable) and one (druggable) is calculated for each pocket. This value represents the closeness to the hyperplane separating the point clouds, indicating how well the point fits into one class. The closer the value is to the two extrema zero and one, the more confident is the prediction.

Descriptor Selection. Before building a model, a subset of meaningful descriptors out of the entire pool of global descriptors generated with DoGSite is selected. A shrinkage discriminant analysis (SDA) is performed for two different classification sets: the NRDD containing all druggable and undruggable pockets and the decoyNRDD including decoys using R.[31] To be more independent from the overall pocket size, descriptors are normalized to the size or occurrence of a certain descriptor; e.g., the number of polar amino acids is divided by the total number of amino acids. This holds true for all descriptors except volume and depth. Nevertheless, very large or too deep pockets should not be overestimated. For this reason, a volume and depth cutoff is introduced. Pockets with a volume larger than 1000 Å^3 or pockets deeper than 30 Å are pruned to these respective values. The 20 amino acid and the five element type descriptors are excluded from the SDA, since they are considered to be too specific for the development of a global model. All used descriptors are provided in Table 3.

First, the pairwise correlation between the remaining 17 descriptors (Table 4) is analyzed. The mean correlation coefficient on the NRDD data set is 0.2. Nevertheless, some descriptors show high correlations. For example, volume, depth, and fitted ellipsoid volume to pocket volume (ellips_vol) have a mean correlation coefficient of 0.7. The same observation can be made for relative polar (aa_pol) and apolar (aa_apol) amino acid number and relative number of donor and acceptors (atm_ac_do). For ranking purposes, we, therefore, incorporate

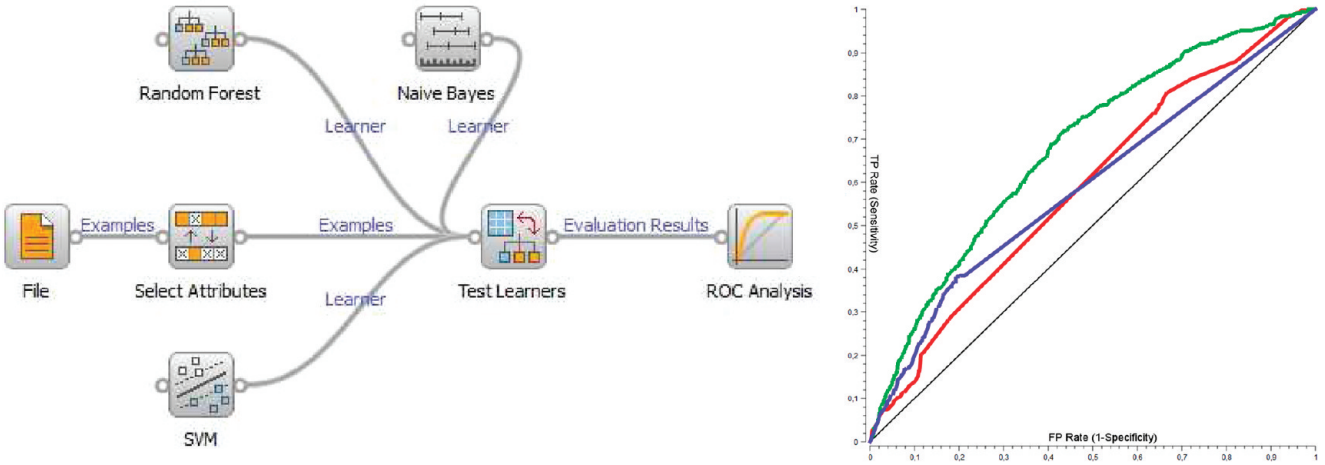


Figure 5. Left: Orange test learner pipeline. Right: Output of Orange ROC analysis, with ROC curves showing performance of SVM (green), BN (blue), and RF (red), respectively.

Table 3. Abbreviations and Explanations of Used Descriptors

descriptor	explanation
volume	volume of the pocket (capped at 1000 Å^3)
depth	depth of the pocket (capped at 30 Å)
ellips_vol	quotient of fitted ellipsoid to pocket volume
ellips_c_a	quotient of every two ellipsoid main axis, with $a>b>c$
ellips_b_a	
ellips_c_b	
gps_se_v	relative number of grid points (gps) of specific types (solvent exposed, surface, volume, and hull)
gps_se_h	
gps_s_v	
aa_apol	relative number of amino acids (aa) of a specific type (apolar, polar, positive, and negative)
aa_pol	
aa_pos	
aa_neg	
atm_ac_do	relative number of acceptor and donor atoms
metals	number of metals present in the pocket
lipo_surf	fraction of lipophilic surface
lipo_siac	relative number of lipophilic SIACs

a shrinkage discriminant analysis accounting for correlations among predictors. Descriptors are ranked according to their ability to separate between the two input classes. Note that in the NRDD, druggable pockets are separated from undruggable ones, while in the decoyNRDD, druggable pockets are separated from undruggable and decoy pockets. High ranking features for the decoyNRDD are dominated by size descriptors (volume, depth, and ellips_vol), while chemical features are more important to separating druggable from undruggable pockets. This justifies the incorporation of two models, one on the decoyNRDD for comparison studies to previously published methods and a second model on the NRDD for druggability studies, e.g., on the DD.

Furthermore, the number of descriptors that is necessary to optimally separate the data, without overfitting the model, was investigated. A model is built on the basis of all 17 descriptors, as well as on the three to nine top ranked descriptors. The model with the highest accuracy and a low number of support vectors is chosen. For the NRDD, best results could be achieved on the basis of the three top ranking descriptors. The decoyNRDD model is built on the basis of the six top ranking descriptors.

Table 4. Feature Ranking Based on a Shrinkage Discriminant Analysis for the Two Input Data Sets NRDD and decoyNRDD

rank	NRDD	decoyNRDD
1	depth	volume
2	aa_apol	depth
3	volume	ellips_vol
4	ellips_c_a	metals
5	lipo_surf	aa_apol
6	aa_pol	lipo_surf
7	atm_ac_do	ellips_c_b
8	ellips_b_a	ellips_c_a
9	ellips_vol	aa_pos
10	aa_pos	lipo_siac
11	lipo_siac	gps_s_v
12	aa_neg	aa_neg
13	gps_se_v	atm_ac_do
14	gps_se_h	aa_pol
15	ellips_c_b	ellips_b_a
16	metals	gps_se_v
17	gps_s_v	gps_se_h

Model Setup. The model is trained on half of the input structures from the decoyNRDD, generated as described in the data section. The remaining half of the data is used for testing purpose. A 50-fold external cross validation is performed. In each run, the data are randomly separated into half training and half test data. During subset creation, it is ensured that both sets contain exactly half of the set of druggable pockets and half of the set of undruggable and decoy pockets. For each model, features are normalized and centered, and a Gaussian kernel is used. Parameters σ of the Gaussian and c of the SVM are set via internal cross validation. The mean accuracy of the predictions on the 50 test sets is 90% for the pocket and subpocket models with a standard deviation below 2%. This shows the stability of the method independent of the chosen 50% test set. The decoyNRDD model is used for enrichment studies and comparison to the previously published methods.

For druggability predictions, a true two-class SVM model is chosen, taking only druggable and undruggable pockets into account and neglecting decoys as well as the intermediate set of difficult targets. The model is trained on the complete NRDD,

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

based on the three top ranking descriptors: depth, fraction of apolar amino acids, and volume.

Simple Score. Besides the classification using the SVM, we are interested in the fact of whether a simple linear combination of discriminant pocket features is able to separate druggable from undruggable pockets. Therefore, a SimpleScore is calculated on the basis of linear regression on the decoyNRDD data set based on three features describing size (volume), enclosure (gps_se_h), and hydrophobicity (lipo_siac) of the pocket. This SimpleScore mimics the SiteMap score introduced by Halgren relying on three similar features:

$$\text{SimpleScore} = -0.62 + 0.035\sqrt{\text{volume}} - 0.016\ \text{gps\_se\_h} + 0.4\ \text{lipo\_siac} \quad (1)$$

B. Local Nearest Neighbor Approach. As a second strategy, the focus on local pocket properties has been pursued. Local functional distance histograms are calculated for all proteins in the data set. These histograms form the basis for a nearest neighbor search. First, a training set with calculated histograms from known druggable and undruggable proteins is collected from the NRDD data set. These histograms form the two template groups H_{drug} and H_{undrug} for the homology-based search. The nearest neighbor score (nnScore) for a query protein is derived in a stepwise manner: For all histograms of the query protein H_{query}, the most similar histogram inside each of the two template groups H_{drug} and H_{undrug} is determined. Given two histograms h_i and h_j, the histogram distance $\text{h\_dist}(h_i,h_j)$ is calculated by summing up the absolute values of prefix sums of the difference in each of the 10 histogram bins.[32]

$$\text{h\_dist}(h_i, h_j) = \sum_{b=0}^{9} \left| \sum_{d=0}^{b} (h_i(d) - h_j(d)) \right| \quad (2)$$

The most similar histogram indicated by the smallest distance value from each group is used for the score calculation of the current query histogram h_q:

$$\text{Score}(h_q) = \min_{\forall h_{\text{ud}} \in H_{\text{undrug}}} \text{h\_dist}(h_q, h_{\text{ud}}) - \min_{\forall h_d \in H_{\text{drug}}} \text{h\_dist}(h_q, h_d) \quad (3)$$

Unspecific histograms obtain a score close to zero, since they are present in both template groups. In the case that a similar histogram is found in the druggable template group and no similar one in the undruggable group, the resulting high score indicates druggability. The final nearest neighbor score nnScore is calculated using the maximal absolute value over the scores of all histograms of one query pocket.

$$\text{nnScore}(H_{\text{query}}) = \text{Score}(h_q) \mid \forall\, h_j \in H_{\text{query}} |\text{Score}(h_q)| > |\text{Score}(h_j)| \quad (4)$$

If the score is below zero, the pocket is considered undruggable. If the score is above zero, it is considered druggable.

RESULTS AND DISCUSSION

Four aspects of DoGSiteScorer are evaluated and discussed. First, the discriminating power of the chosen descriptors is investigated. Subsequently, the global SVM model is evaluated and compared to previous publications on the decoyNRDD data set, and predictions are made for the DD on the basis of the NRDD model. Third, the functional group histograms and the local nearest neighbor method are analyzed with respect to their ability to discriminate druggable from undruggable pockets. Finally, the correlation between the two methods is investigated, and scope and limitations of the druggability prediction are highlighted.

Descriptor Analysis. As described in the methods section, a small set of descriptors has been chosen to distinguish between druggable and undruggable pockets. The descriptors are calculated for all DD pockets separated into three classes: druggable, undruggable, and difficult (see Figure 6 for corresponding boxplots). Our findings for global pocket descriptors agree with properties already described as being important in the separation of druggable and undruggable pockets.[1,10,11,17,21] Druggable pockets tend to be larger, more hydrophilic, and complex in shape. The average volumes of 900 Å^3 for druggable pockets and 300 Å^3 for undruggable pockets highly agree with values reviewed by Egner and Hillig in 2008 of 930 Å^3 and 330 Å^3 for the respective pocket classes.[1] The depth of a pocket is another distinguishing factor. Druggable pockets are deeper, with average depth values of 21 Å compared to 13 Å of undruggable pockets. While the ellipsoidal main axis ratio ellips_c_a was ranked high by the discriminant analysis on the NRDD, it seems to have moderate discrimination power on the DD with alike mean values for all three classes. On the contrary, the lower ranked description of pockets enclosure (gps_se_h) allows for a clear separation between the classes. Undruggable pockets are more solvent exposed, indicated by the higher solvent exposed to hull grid points ratio of 0.17 compared to druggable pockets with 0.08. Besides being larger and more enclosed, a higher apolar character of the pockets is another indicator for druggability. The fraction of apolar amino acids proved to be a highly discriminative property with mean values of 0.57 and 0.37 for druggable and undruggable pockets, respectively. Likewise, the mean lipophilic surface fraction value for druggable pockets lies at 0.7 for druggable and 0.6 for undruggable pockets, concurring with recently published values.[1,21] The higher lipophilic character of druggable pockets is moreover exposed in a lower fraction of donor and acceptor atoms in the pocket. When considering pockets from the difficult category, we found that their size and shape properties closer resemble druggable pockets; e.g., their average volume is 820 Å^3, the average depth 20 Å, and the pocket enclosure 0.1. In terms of physicochemical properties, nevertheless, they approach undruggable pockets, showing the same average apolar surface fraction of 0.6 as undruggable pockets. This observation holds true for other chemical pocket features, e.g., the fraction of apolar amino acids. While evaluating SiteMap,[17] Halgren made a similar finding, stating that size and enclosure are better suitable for distinguishing between undruggable and difficult sites, while hydrophobicity helps to separate difficult from druggable pockets.

SVM Performance. To evaluate our method, descriptor-based enrichment studies are performed on the decoyNRDD, prepared as described in the methods section. In this test scenario, predictions are made for the complete decoyNRDD data set and the pockets are sorted by their druggability score. A normalized enrichment factor is used, as introduced by Schmidtke and Barril, measuring the ratio of druggable pockets in the given percentage subset of the data.[21] The percentage of selected data is plotted against the relative enrichment. Pocket and subpocket decoyNRDD models based on the six chosen

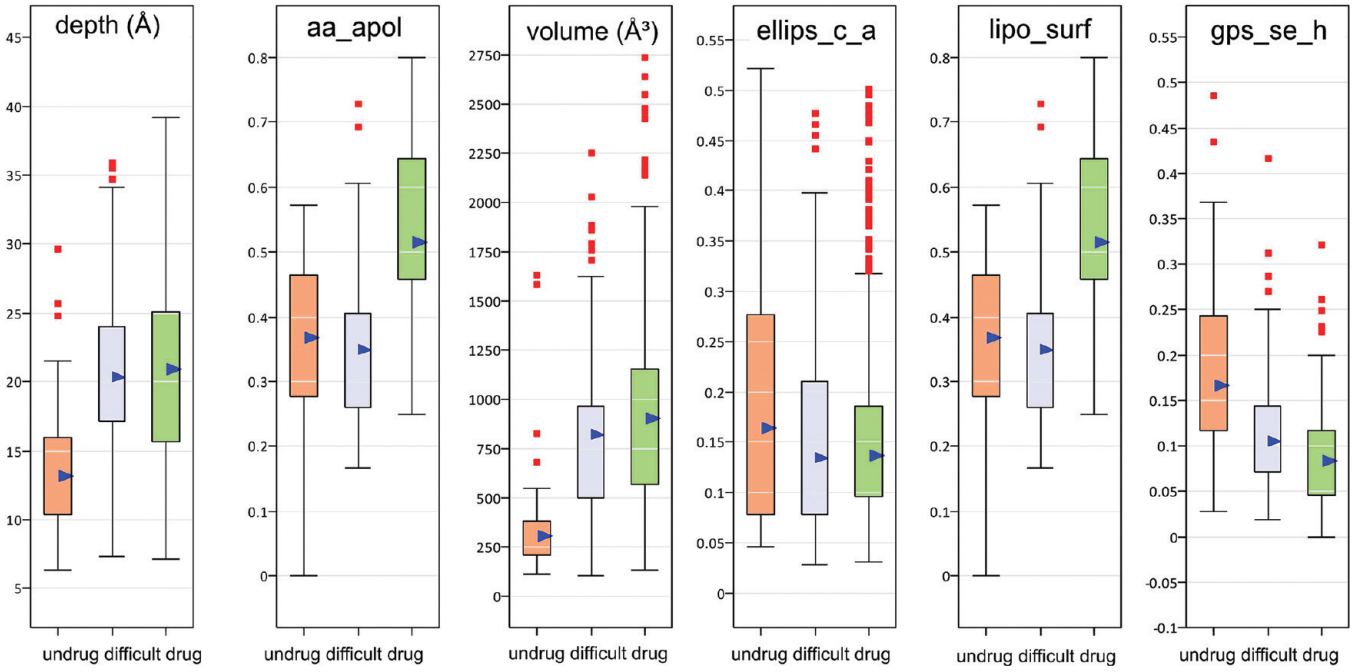


Figure 6. A boxplot is exemplarily shown for six global descriptors. From left to right: depth, fraction of apolar amino acids, volume, ratio between ellipsoid main axes c and a, lipophilic surface fraction, and ratio of the solvent exposed to hull grid points. Boxes are shown for undruggable, difficult, and druggable pockets, separated as shown in Figure 1. Shown in the boxplot are the median value (blue triangle), the upper and lower quartile (box), the upper and lower adjacent values (horizontal black lines), and the outliers (red squares).

descriptors are compared to the results of the recently published algorithms fpocket and SiteMap (Figure 7). In addition to the

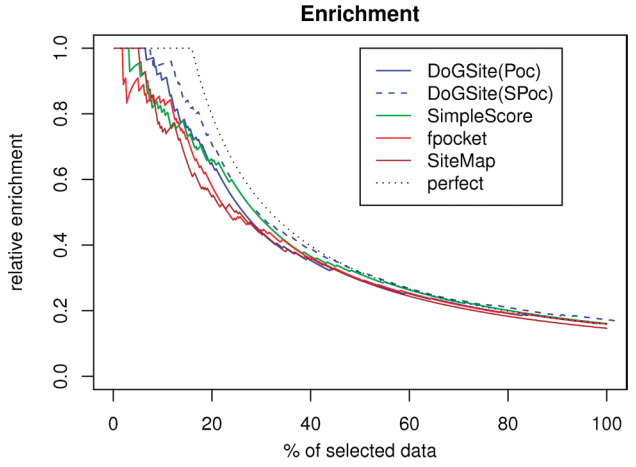


Figure 7. Comparison of DoGSiteScorer to two other druggability prediction methods (data for fpocket and SiteMap were provided by Schmidtke and Barril).

individual druggability scores, the SimpleScore is plotted. The gray curve denotes optimal performance. As already discussed in the Material and Methods section, the number of considered pockets constituting 100% in the enrichment analysis differs slightly for each method (see Table 2). All three methods have high prediction accuracy. Nevertheless, DoGSiteScorer (blue lines) shows the highest performance, especially when considering subpockets. Furthermore, the introduced SimpleScore performs quite similarly to SiteMaps DScore, which is not surprising since both scores are linear combinations based on three similar descriptors coding size, complexity, and hydrophobicity of the pocket.

As mentioned in the Material and Methods section, using decoys to enrich the negative data source introduces an unwanted structural bias. Pockets are rather separated by size than by their overall ability to bind a bioavailable drug molecule. Incorporating decoys when training the model blends the terms druggability and ligandability. The negative set includes pockets binding a molecule that is not druglike or not bioavailable and pockets that are empty. For those decoy pockets it is not proven whether either the pocket is not able to bind a ligand or a ligand exists and is missing in the present crystal structure.

Thus, in the second model, only druggable and undruggable pockets from the NRDD are incorporated, containing 89 pockets in total. Training the support vector machine on such a small data set bears the risk of overfitting. An analysis showed that using the three top ranking descriptors' volume, fraction of apolar amino acids, and depth yields the best results. We are aware of the fact that the nature of druggable pockets is more complex and would more consistently be described with an increasing number of descriptors and by models based on larger trainings data sets. This underlines especially the need of additional negative data. However, high prediction accuracies show that these three global descriptors inherit necessary features to distinguish druggable from undruggable targets to a decent amount.

DoGSiteScorer is a fully automated method able to perform high throughput screens. Hence, the performance and the stability of the NRDD druggability model are evaluated on the complete DD. Targets are grouped by functional class, and a boxplot of resulting class dependent scores is shown in Figure 8, for druggable, difficult, and undruggable targets, separately. The box widths reflect the number of structures present in each group. Druggability scores range from zero to one. The druggability cutoff is set to 0.5, indicated by a blue horizontal line. Targets with scores above that line are considered as being druggable. Regarding the average druggability score per class

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

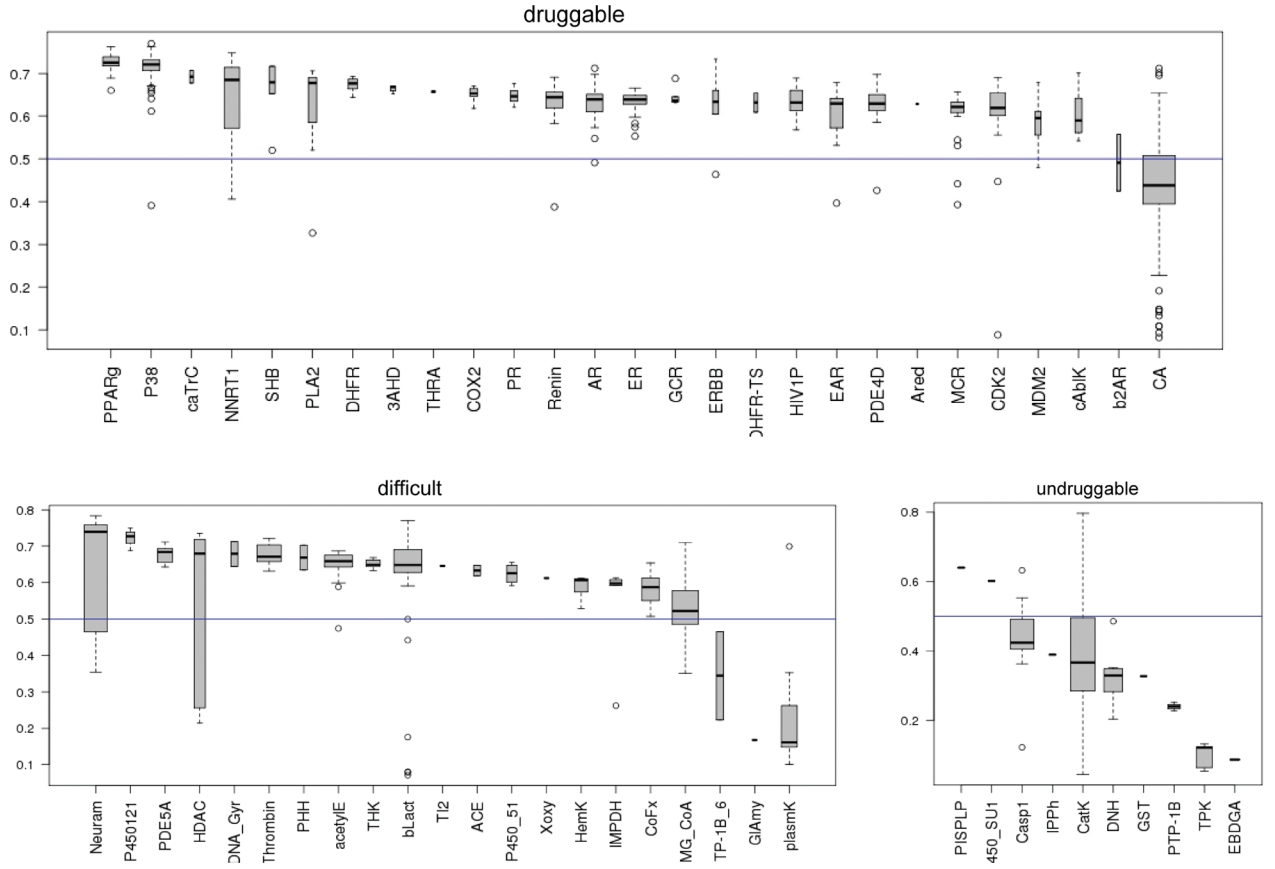


Figure 8. A boxplot of druggability scores predicted on the DD is shown. Structures are grouped by their functional class. Mean druggability values per class are indicated by the solid black lines in the boxes. The blue line indicates the 0.5 druggability cutoff.

(solid black line in the respective box), 88% of all target families are correctly predicted. Predictions made on the DD by DoGSiteScorer are in high accordance with the real classification of the data.

Common drawbacks are problems, particularly during pocket predictions, with shallow or very flexible sites. If the drug binding pocket is not found by the pocket detection algorithm, or if it does not fulfill the required coverage criterion, it is not listed in Figure 8. For example, the shallow binding sites of HIV-integrase and HIV-RT are correctly predicted neither by DoGSiteScorer nor by fpocket. Some further target classes (e.g., cycloC, ABT, and ADAM33) which are present in the fpocket study cannot be considered in this druggability evaluation due to the coverage criterion in pocket prediction. On the contrary, DoGSite predicts, e.g., the correct pocket for coagulation factor x (CoFx), while it is split iton two parts by fpocket. For this reason, fpocket ranks CoFx poorly (score around 0.2), while DoGSiteScorer predicts an average score of 0.6 and therefore a druggable pocket. Predictions are very robust for structures with well defined binding pockets. For these structures, pockets can be confidently predicted and are not influenced by small residual changes. This holds especially true for targets such as most nuclear hormone receptors: peroxisome proliferator-activated receptor gamma (PPARg), progesteron (PR), androgen receptor (AR), estrogen (ER), glutocorticoid (GCR), and mineralcorticoid (MCR). These hormone receptors have a clearly defined deep pocket, and the scored descriptors fit well into the druggability ranges. For example, values for estrogen receptor structure 1err are a volume of around 1000 Å^3, an apolar amino acid fraction of 0.7, and a depth of 25 Å. MAP p38α protein kinases are an example for a beautiful binding site (Figure 9) and the robustness of the

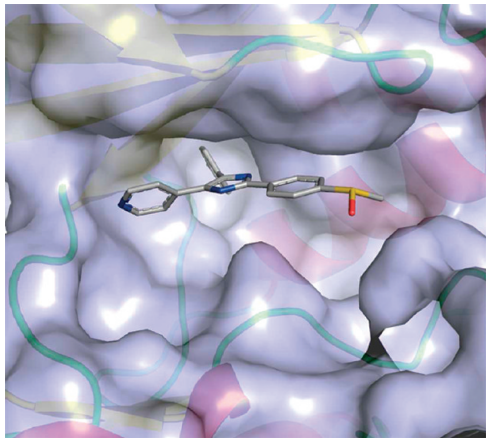


Figure 9. Binding site of MAP p38α binding site (PDB code 1a9u) with bound imidazoyl-based inhibitor. The protein contains a well-formed binding site.

predictor. There are 40 different p38 structures present in the analysis. Although the considered pockets span a large volume range from 450 Å^3 to almost 1800 Å^3, they are correctly classified as druggable. The reason for the observed volume range is that the structures are crystallized in different activation states. Most notably, p38α structures in the DFG-in and DFG-out conformations are present in the data set. The p38α

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

complex structure with the smallest volume is a DFG-in structure in complex with a quinoline-based inhibitor (PDB code 2zaz, druggability score 0.65). The p38α in the DFG-out conformation (PDB code 1wbs, druggability score 0.74) has the largest volume. As an important descriptor for druggability, all p38α pockets exhibit a high fraction of lipophilic surface.

In contrast, four druggable and difficult target classes cannot be predicted correctly, namely, carbonic anhydrases (CA), plasminogen kringle 4 (plasmK), glucoamylases (GlAmy), and protein tyrosine phosphatases (PTP-1B). PlasmK and PTP-1B are similarly predicted as being undruggable in the fpocket study. For PTP-1B, e.g., it was claimed that the allosteric binding sites of these structures hold only weak binders; hence the annotation of lower druggability scores is comprehensible. Glucoamylases bind an oral drug and are listed in the DrugBank. However, the drug is not bioavailable and acts in the intestine. In the studies of Schmidtke and Halgren, these structures were originally introduced into the undruggable data sets. Nevertheless, both methods fpocket and SiteMap predicted them as being druggable, while DoGSiteScorer annotates the pockets as undruggable. This example shows that in some cases a clear assignment of druggability can be very difficult.

Carbonic anhydrase targets constitute the only class from the druggability set with a mean value below 0.5. Carbonic anhydrases have two characteristics which complicate a correct prediction. First, some structures (e.g., PDB codes 2qoa and 2nns) have multiple inhibitor molecules bound (Figure 10).

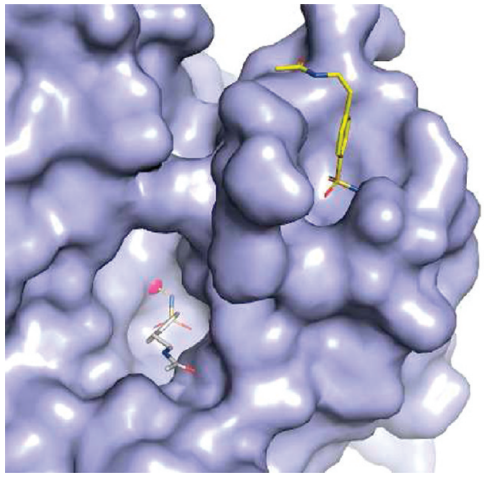


Figure 10. Carbonic anhydrase with two sulfonamide ligands bound. One binds to the catalytic binding pocket and coordinates to the catalytic zinc and histidines (gray carbons), whereas another sulfonamide inhibitor compound is bound to a small binding pocket located on the protein surface (carbons in yellow).

Whereas one inhibitor molecule binds to the known catalytic site near the zinc atom and the three catalytic histidines, another ligand molecule is bound on a surface cavity which is not involved in the catalytic function of the enzyme. This surface pocket is correctly predicted as undruggable. However since the real "drug-like" inhibitor is bound, it tampers the druggability prediction results. Second, carbonic anhydrase pockets are rather small and hydrophilic. Furthermore, the ligand binds via metal interactions, which is not covered as a descriptor in this analysis. These features do not match the required known global druggability features, describing a large, complex, and hydrophobic pocket. The same problem is encountered for HMG_CoA reductase structures which obtain a borderline mean druggability score of 0.5. The binding of the drug is dominated by hydrogen bonds and ionic interactions. Druggability results in such structures from single interactions rather than from global pocket features, which motivates a more local homology-driven approach, as discussed in the next section.

Druggability scores are predicted on apo and holo structures. The generally low standard deviation inside the individual classes shows the robustness of the method. Anyhow, some examples exist where high standard deviations are encountered. The protein family with the largest deviation in their druggability scores is the class of histone deacetylases (HDAC8, see Figure 11). The family

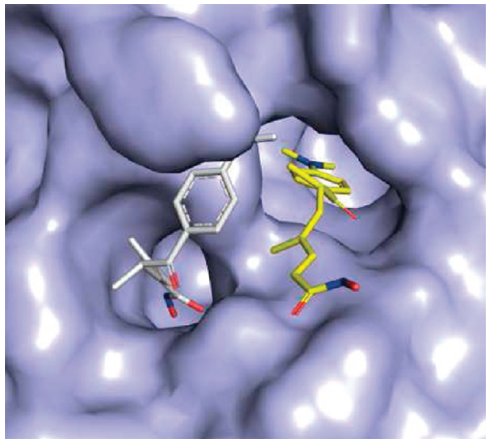


Figure 11. The binding site of an HDAC8 structure (PDB code 1t64) is shown. Two hydroxamate inhibitors are bound.The left molecule (carbons in gray) binds to the catalytic site and interacts via the hydroxamate to the catalytic histidines 142 and 143. The second bound inhibitor molecule (carbons in yellow) binds to an adjacent pocket with the hydroxamate pointing outside the cavity.

comprises three different structures, where two entries (PDB codes 1w22 and 1t64) consist of a homodimer with two to four inhibitors bound and the last entry represents a monomer (PDB code 1t67). In the case of the 1w22 structure, both pockets are predicted to be druggable (druggability scores 0.74 and 0.72). In the case of structure 1t64 also, a homodimer is found; however, in each binding site, there are two inhibitor molecules found. The pocket which is important for the enzymatic reactions that contains the catalytic histidines is found entirely and predicted as druggable. However, the second pocket is not found as one entire pocket. Therefore, e.g., the volumes differ, and the second adjacent pocket is too small to be considered as druggable.

As exemplarily revealed in some of the case studies, wrong or misleading data strongly affect the prediction power of the method. This is a pitfall in the high throughput druggability prediction field since dubious prediction results have to be manually analyzed.

Nearest Neighbor Performance. In order to facilitate a nearest neighbor based classification, the NRDD pockets form the training set as in the SVM-based druggability prediction experiment. All histograms from druggable and undruggable NRDD pockets are collected and constitute the two template groups as described in eq 3. Druggable and undruggable pockets are compared with respect to their binding motifs. All unspecific histograms that are present in both template classes,

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

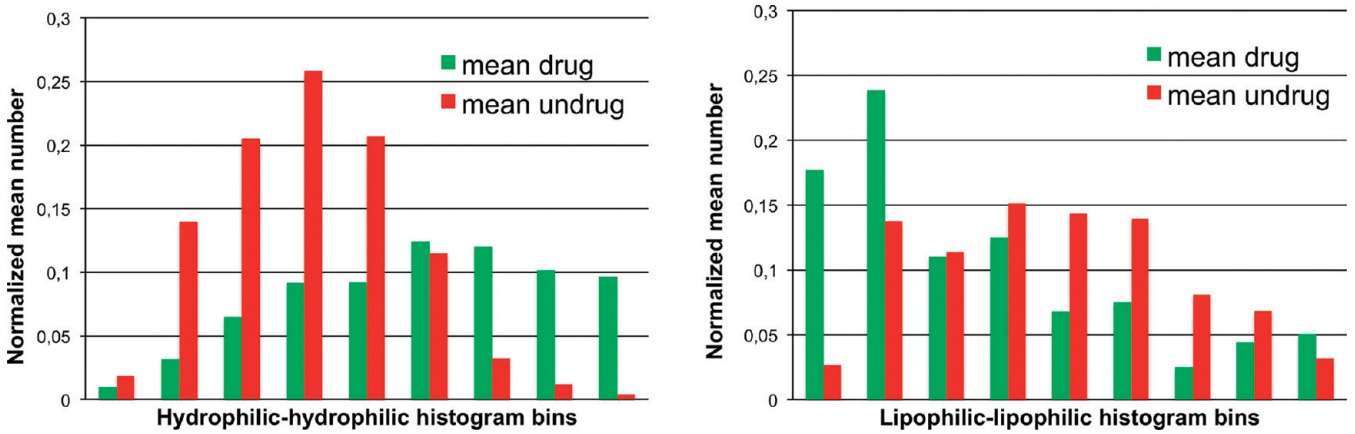


Figure 12. Comparison of druggable and undruggable target binding motifs for hydrophilic−hydrophilic (left) and lipophilic−lipophilic distance pairs (right). On the *x* axis, the 2 Å distance bins are plotted; on the *y* axis, the normalized mean number of pairs over the respective data class. The histograms are individually normalized such that the area under the curves is 1.0.

yielding a score close to zero, are neglected. The remaining histograms are compared on the basis of hydrophilic−hydrophilic and lipophilic−lipophilic atom pairs, respectively. As shown in Figure 12, undruggable targets tend to have more short-range hydrophilic−hydrophilic interactions and less short-range lipophilic−lipophilic interactions.

Subsequently, the DD pockets are used as queries, and nnScores are predicted on the basis of hydrophilic−hydrophilic histograms. Note that although the NRDD is a subset of the DD, it is guaranteed that two histograms originating from the same structure are never compared during the nearest neighbor search. In 88% of the cases, the algorithm predicts the correct druggability state.

Boosting Performance by the Nearest Neighbor Search. Both global and local methods can be used in an automatic manner. Nevertheless, we decided not to directly incorporate the local nearest neighbor value in the SVM to avoid overfitting by learning on a learned value. Both methods comprise individual information based on global and local features. The information gain of incorporating the second method is shown in Figure 13. Targets are again grouped into

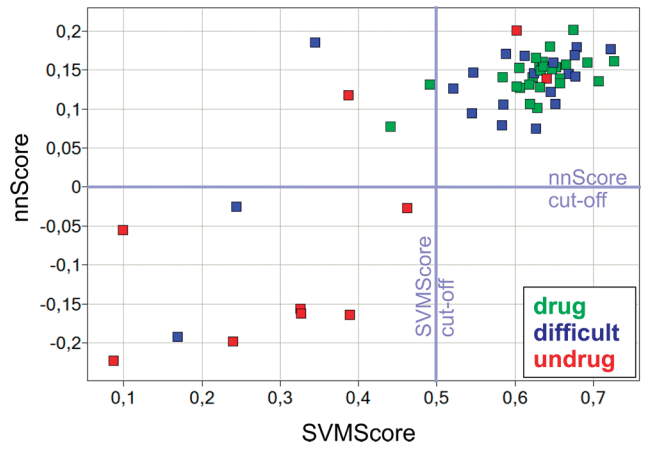


Figure 13. Comparison of predicted SVMScore (*x* axis) against nnScore (*y* axis) on the DD target classes. Respective score cut-offs are shown in blue. Data point colors indicate known druggability state.

functional classes, and the mean value per class is calculated. The mean SVMScore is plotted against the mean nnScore of the respective target class. Note that the SVM druggability cutoff lies at 0.5, while the nnScore cutoff lies at 0.0. As indicated by color in Figure 13, most druggable pockets are predicted as druggable by both methods (upper right corner), as well as most undruggable pockets as undruggable (lower left corner).

Most interesting are the cases where the two methods disagree. For example, there are cases where the global SVM method predicts low scores for druggable pockets, while the local nearest neighbor method yields higher druggability scores. In this context, carbonic anhydrases are further discussed since they failed in the global assignment (mean SVMScore: 0.44) but are correctly predicted by the local method (mean nnScore: 0.08). The failure is mostly due to the globally undruggable character of the pocket. For 40% of all carbonic anhydrase pockets, the carbonic anhydrase present in the template set (3caj) was found as the nearest neighbor. Nevertheless, when considering the histograms that are responsible for the final nnScore, a carbonic anhydrase specific binding motif is found for most pockets. Exemplarily, four nearest neighbor histograms as well as the mean value over all nearest neighbor histograms of the carbonic anhydrase class are registered in Table 5.

Table 5. Typical Histograms Found during the Nearest Neighbor Search for Carbonic Anhydrase Pockets[a]

	hydrophilic−hydrophilic atom pair distance bins in Å								
protein	0−2	2−4	4−6	6−8	8−10	10−12	12−14	14−16	16−18
2nno	4	0	4	5	2	0	0	0	0
1caz	2	2	2	6	3	0	0	0	0
2q38	3	1	2	6	3	0	0	0	0
3caj	2	1	3	7	2	0	0	0	0
mean	2.7	1	2.6	6.5	2	0.4	0	0	0

[a]First column denotes PDB codes of the proteins; subsequent columns describe the distance bins in 2 Å steps for hydrophilic−hydrophilic atom pairs.

On average, 15 functional atom partners are considered in the histograms representing the hydrophilic character of these pockets. Furthermore, very few large distance partners are found, reflecting the small volume of the pocket. In Figure 14, the binding site of carbonic anhydrase structure 3caj with bound ligands is shown. The found histograms represent the interactions necessary for ligand binding (Figure 14). Depending on the atom from which the histogram originates, approxi-

dx.doi.org/10.1021/ci200454v | J. Chem. Inf. Model. 2012, 52, 360−372

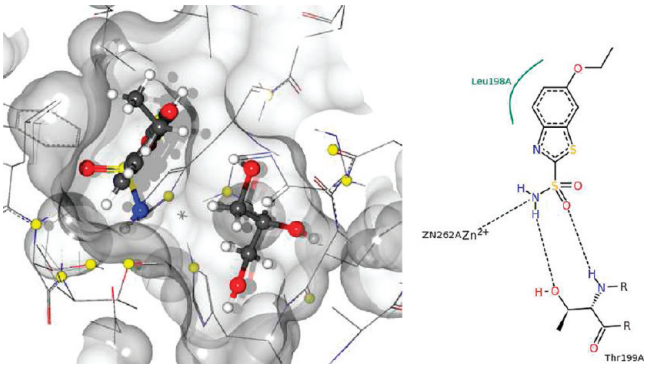


Figure 14. Binding site of carbonic anhydrase structure with bound ligands (EZL, GOL). Additionally, hydrophilic atoms considered in the histogram calculation are marked by yellow spheres. On the right side, the ligand binding motif of EZL is shown (drawn with Poseview).

mately three close partners are found, representing the three catalytic histidines, coordinating the zinc ion. Besides interaction with the ion, the ligand binds to threonine 199, which is indicated by the next farther partners in the histogram. The functional atoms farther away are probably incorporated in binding the glycerol molecule.

In conclusion, both methods, the nearest neighbor and the SVM method, have high prediction accuracies. The above example showed the boosting power of the nearest neighbor search by exhibiting special binding features of a pocket. But again, the need of a broad underlying template data set should be mentioned, since the nearest neighbor method is generally based on a homology search.

CONCLUSION

In this work, we present DoGSiteScorer, a new algorithm for fully automatic druggability predictions starting from the three-dimensional structure of the protein. Three to six global descriptors are used to train a SVM model, with which druggability scores can be predicted for new targets. These descriptors represent known druggable features encoding for size, compactness, and physicochemical properties of the pockets. Enrichment studies on the decoyNRDD show that our method performs comparably well to SiteMap and fpocket, with the subpocket-based model yielding the field. Furthermore, the algorithm correctly classifies 88% of the DD targets into the accurate druggability class, underlining the power of the use of global descriptors. Nevertheless, classifications fail in the case of pockets which do not hold these global features. In these cases, single interactions are the key for druggability, rather than the global size or shape of the pocket. Therefore, we introduced a nearest neighbor search based on distance dependent histograms. Analysis of the distances between individual points in the pocket show significant differences between druggable and undruggable pockets. We found that druggable pockets tend to have less short-range hydrophilic interaction pairs and more short-range lipophilic pairs compared to undruggable pockets. Tested on the DD, the nearest neighbor search identified the correct druggability class in 88% of the data. Local and global predictors can be used in an automatic manner and are suitable for high throughput screens. Combining both values increases the reliability of druggability predictions. Nevertheless, some challenges remain. A drawback of fully automated methods is uncertainties due to the pocket prediction step. Unspecific or wrong pockets may mislead the subsequent druggability predictions. Flexibility of the pocket upon ligand binding adds to the complexity of the problem. Furthermore, the lack of negative data and the ambiguity in the definition of the term druggability make training of high throughput methods difficult. Nevertheless, methods like DoGSiteScorer provide qualitatively and quantitatively valuable data for drug target assessment. In particular, the identification and rating of allosteric sites can shed light onto new pockets for the drug discovery process.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

ACKNOWLEDGMENTS

The project is part of the Biokatalyse2021 cluster and funded by the BMBF under grant 0315292A. We would like to thank Peter Schmidtke and Xavier Barril for collecting the NRDD and DD data sets and making them available to the community. Without their effort, this study would not have been possible. Furthermore, we would like to thank our cooperation partner BioSolveIT GmbH, especially C. Lemmen, for helpful discussions.

REFERENCES

(1) Egner, U.; Hillig, R. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391−401.
(2) Hopkins, A.; Groom, C. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.
(3) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067−1077.
(4) Hopkins, A.; Groom, C. Target analysis: a priori assessment of druggability. *Ernst Schering Res. Found. Workshop* **2003**, *11*, 17.
(5) Hajduk, P.; Huth, J.; Tse, C. Predicting protein druggability. *Drug Discovery Today* **2005**, *10*, 1675−1682.
(6) Sakharkar, M.; Sakharkar, K.; Pervaiz, S. Druggability of human disease genes. *Int. J. Biochem. Cell B.* **2007**, *39*, 1156−1164.
(7) Sheridan, R.; Maiorov, V.; Holloway, M.; Cornell, W.; Gao, Y.-D. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J. Chem. Inf. Model.* **2010**, *50*, 2029−2040.
(8) Edfeldt, F.; Folmer, R.; Breeze, A. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today* **2011**, *16*, 284−287.
(9) Ward, R. Using protein-ligand docking to assess the chemical tractability of inhibiting a protein target. *J. Mol. Model.* **2010**, *16*, 1833−1843.
(10) Hajduk, P.; Huth, J.; Fesik, S. Druggability indices for protein targets derived from NMRbased screening data. *J. Med. Chem.* **2005**, *48*, 2518−2525.
(11) Cheng, A.; Coleman, R.; Smyth, K.; Cao, Q.; Soulard, P.; Caffrey, D.; Salzberg, A.; Huang, E. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71−75.
(12) Dalvit, C. NMR methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug Discovery Today* **2009**, *14*, 1051−1057.
(13) Zheng, C.; Han, L.; Yap, C.; Ji, Z.; Cao, Z.; Chen, Y. Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* **2006**, *58*, 259−279.
(14) Zheng, C.; Han, L.; Yap, C.; Xie, B.; Chen, Y. Progress and problems in the exploration of therapeutic targets. *Drug Discovery Today* **2006**, *11*, 412−420.
(15) Han, L.; Zheng, C.; Xie, B.; Jia, J.; Ma, X.; Zhu, F.; Lin, H.; Chen, X.; Chen, Y. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discovery Today* **2007**, *12*, 304−313.

(16) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identication of drug-binding sites. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 892−906.
(17) Halgren, T. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, 377−389.
(18) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, *10*, 168.
(19) Laurie, A.; Jackson, R. Methods for the prediction of protein-ligand binding sites forstructure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.* **2006**, *7*, 395−406.
(20) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.
(21) Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.
(22) Vieth, M.; Siegel, M.; Higgs, R.; Watson, I.; Robertson, D.; Savin, K.; Durst, G.; Hipskind, P. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224−232.
(23) Wishart, D.; Knox, C.; Guo, A.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−72.
(24) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.
(25) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding af_nities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(26) Schellhammer, I.; Rarey, M. FlexX-Scan: fast, structure-based virtual screening. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 504−517.
(27) Schellhammer, I.; Rarey, M. TrixX: structure-based molecule indexing for large-scale virtual screening in sublinear time. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 223−238.
(28) Rarey, M.; We_ng, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41−54.
(29) Demsar, J.; Zupan, B.; Leban, G.; Curk, T. In *Knowledge Discovery in Databases: PKDD 2004*; Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D., Eds.; Springer: Berlin, 2004; Lecture Notes in Computer Science, vol. *3202*; pp 537−539. http://orange.biolab.si/ (accessed March 1, 2011).
(30) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2011**, *2*, 27:1−27:27. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed August 4, 2010).
(31) Ahdesmaki, M.; Strimmer, K. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.* **2010**, *4*, 503−519. R package version 1.2.0, retrieved from http://CRAN.R-project.org/package=sda (accessed May 4, 2011).
(32) Cha, S.; Srihari, S. On measuring the distance between histograms. *Pattern Recognit.* **2002**, *35*, 1355−1370.