JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

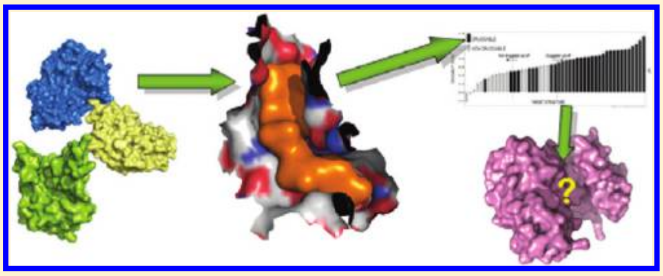
pubs.acs.org/jcim

DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set

Agata Krasowski,[†] Daniel Muthas,[†] Aurijit Sarkar,[†] Stefan Schmitt,[‡] and Ruth Brenk*,[†]

[†]College of Life Sciences, Division of Biological Chemistry and Drug Discovery, University of Dundee, Dow St, Dundee DD1 5EH, U.K.
[‡]AstraZeneca R&D Mölndal, Pepparedsleden 1, S-43184 Mölndal, Sweden

Supporting Information



ABSTRACT: Judging if a protein is able to bind orally available molecules with high affinity, i.e. if a protein is druggable, is an important step in target assessment. In order to derive a structure-based method to predict protein druggability, a comprehensive, nonredundant data set containing crystal structures of 71 druggable and 44 less druggable proteins was compiled by literature search and data mining. This data set was subsequently used to train a structure-based druggability predictor (DrugPred) using partial least-squares projection to latent structures discriminant analysis (PLS-DA). DrugPred performed well in discriminating druggable from less druggable binding sites for both internal and external predictions. The method is robust against conformational changes in the binding site and outperforms previously published methods. The superior performance of DrugPred is likely due to the size and composition of the training set which, in contrast to most previously developed methods, only contains cavities that have evolved to bind a natural ligand.

INTRODUCTION

Sequencing of the human genome and that of many pathogens followed by structural genomics initiatives sparked the expectation that this knowledge would lead to a wealth of new drugs.[1−5] Unfortunately, these advances have not yet translated into an increased success rate in drug discovery. This is perhaps a consequence of the fact that it is not possible to develop a drug, or even a lead, for every protein target.[6] The ability of a target to bind small, drug-like molecules with a high affinity is sometimes referred to as druggability[7,8] and is a very important criteria in the target selection phase.[9] Orally available small molecules remain the favored route of administration. Proteins that can be modulated by such molecules are therefore preferred targets for drug discovery. It is known that passively absorbed, orally available compounds occupy a limited area of chemical space.[10−14] Accordingly, it is reasonable to assume that binding sites which accommodate such molecules, i.e. that are druggable, share some common features.

While commonly used in literature, the term "druggable" is prone to confusion. In the context of target assessment, druggability only refers to the ability of a binding site to bind a drug-like ligand.[7,8] Other critical aspects for drug discovery such as the chemical tractability of the developed compound series, selectivity, and efficacy of the compounds in relevant models are omitted. To reflect this, some researchers have suggested using the term "bindability" instead of "druggability" when describing the properties of binding site pockets with respect to the type of compounds they can potentially accommodate.[15] Since the latter term is more common in literature and also refers to the properties of the compounds that potentially bind to the receptor (e.g., that they are drug-like), we opted to use this expression throughout the text.

Several studies have focused on developing *in silico* methods to predict the druggability of a potential drug target based on its 3D structure. An important early contribution to this field came from Hajduk et al.[16] who used hit rates from fragment screening as a measure of protein druggability. This measure was taken as the dependent variable in regression analysis to derive a druggability predictor. The resulting model was based on geometrical and physicochemical descriptors like polar and apolar surface area, surface complexity, and pocket dimensions. Cheng and coworkers showed that an approximation of the hydrophobic desolvation energy of the protein binding site can be used to predict the maximal affinity achievable between a target and an orally available compound.[17] The derived MAP_{POD} score could then be used to distinguish druggable from undruggable and difficult binding sites. For the same purpose, Halgren developed Dscore as part of the SiteMap package.[18] This scoring function incorporates terms for the size of the binding pocket, its enclosure, and a penalty for its hydrophilicity. Soga and colleagues investigated how the amino acid composition in the binding site varied as compared to the protein surface in general.[19] They showed that the derived propensity for ligand binding (PLB) score was a useful metric for identifying true binding sites and

Received: June 14, 2011
Published: October 13, 2011

Table 1. Composition of Data Sets Used to Derive and Test Structure-Based Druggability Prediction Methods

	Hajduk set[16]	Cheng set[a,17]	DD set[b,20]	NRDD set[b,20]
number of druggable binding sites	35	43 (17)	919	45
number of less druggable binding sites	37	10 (4)	84	20
number of prodrug sites		10 (5)	67	5
total number of binding sites	72	63 (27)	1070	70

[a] Number of unique binding sites is given in brackets. [b] The NRDD set is the nonredundant subset of the DD set.

especially for those harboring drug-like molecules. Schmidtke and Barril published a druggability predictor derived by logistic regression analysis.[20] This predictor is based on descriptors capturing the hydrophobicity and polarity of the binding site. Recently, Sheridan et al. proposed a method to predict the "bindability" of a binding site (defined as being able to bind drug-like molecules) based on its local neighborhood in pocket space to binding sites that contain drug-like cocrystallized ligands.[15] In their metric which is referred to as drug-like density (DLID), pocket space is defined by the three parameters volume, buriedness, and hydrophobicity.

Data sets of different size and composition suitable to generate structure-based druggability prediction methods have been published (Table 1). Hajduk et al. made a data set containing 72 unique protein binding sites (here referred to as the Hajduk set) publically available.[16] Cheng et al. compiled a data set compromising 63 structures out of which 27 are unique (here referred to as the Cheng set).[17] Schmidtke and Barril published a large but redundant set of 1070 structures (called DD set) and a nonredundant subset (called NRDD set) containing 70 binding sites. Finally, Sheridan et al. derived a data set by calculating surface cavities for all structures stored in the PDB with a resolution ≤3 Å. They ended up with ~290,000 pockets out of which ~5700 contained a drug-like ligand.

Here, we introduce a new structure-based druggability predictor (DrugPred). In this context, a binding site was defined to be druggable if it can noncovalently bind small, drug-like ligands which are orally available and do not require administration as prodrugs. While covalently binding drugs and prodrugs constitute an important group of clinically used molecules, they were excluded by this definition as being druggable to enable deriving a homogeneous data set for training and testing of the predictor.[21,22] In the case of covalent drugs, a significant amount of the binding energy originates from the covalent bond that the drug molecule forms with its target. Prodrug strategies are often applied when otherwise orally available ligands can not be developed. Therefore, in both cases the binding site characteristics are likely to differ from pockets binding noncovalently orally available molecules. This assumption is supported by previously developed druggability methods which score such binding sites in the same region as nondruggable or undruggable pockets.[17,20] Consequently, for model building targets for which only prodrugs or covalently binding molecules are known should not be grouped together with other druggable targets. Further, it has to be kept in mind that druggability is not an absolute measure. With a lot of effort and resources orally available drugs targeting challenging binding sites might eventually be developed, as illustrated by compounds targeting factor Xa.[23] To reflect this fact, we named binding sites for which druggability has not yet been shown "less druggable".

DrugPred was derived using partial least-squares projection to latent structures discriminant analysis (PLS-DA).[24] To facilitate this task, we compiled an extensive data set containing the crystal structures of 71 druggable and 44 less druggable protein binding sites. To the best of our knowledge, this is the largest nonredundant data set for this purpose publically available to date. We show that DrugPred has a good predictive power in discriminating druggable from less druggable binding sites, both in terms of internal and external predictions and outperforms most previously published methods. Further, we demonstrate that DrugPred is robust against conformational changes in the binding sites. We also discuss possible reasons for the superior performance of DrugPred compared to other druggability prediction methods.

RESULTS

Data Set of Druggable and Less Druggable Protein Binding Sites. An extensive, non-redundant set of druggable and less druggable binding sites (NRDLD set) was compiled in order to facilitate statistical modeling and to enable us to develop a druggability predictor. Crystal structures of druggable binding sites were identified by consulting the relevant literature[25−27] and DrugBank[28] for targets of orally available, noncovalent drugs that do not require uptake as prodrugs. These structures were augmented with the druggable proteins published by Cheng et al.[17] and the structures from the Astex diverse set, which contains crystal structures in complex with drug-like ligands.[29] A sequence alignment was performed for enzymes having the same EC number, and structures with a sequence identity higher than 60% to other members in the same group were filtered out in order to avoid redundancy. This resulted in a set of 71 druggable binding sites (Table S1, Supporting Information).

Only limited information about experimentally confirmed less druggable targets is publically available, and to the best of our knowledge, only one comprehensive review has been published.[6] Out of over 70 targets in this review, the crystal structures for only two proteins have been deposited in the PDB. For training a druggability predictor, a larger set is necessary. Additional less druggable binding sites were therefore identified by mining the data stored in PDBbind,[30] DrugBank,[28] and ChEMBL.[31] In agreement with our definition of druggable proteins, we postulated that a protein is less druggable if none of the following requirements are met: 1) At least one ligand is orally available as judged by Lipinski's "rule-of-five".[11] 2) In addition, these ligands must have a clogP ≥ −2. While Lipinski's rule clearly places an upper limit on the clogP value for drug-like molecules, it has also been established that overtly hydrophilic compounds are poorly absorbed.[14] Hence a low clogP was also used to predict poor bioavailability. 3) Further, the ligand efficiency of at least one of the ligands fulfilling criteria 1 and 2 is ≥0.3 kcal mol^{-1}/heavy atom.[32] Drugs usually have affinities in the low nanomolar range for their targets. If they adhere to Lipinski's "rule-of-five" this translates to a ligand efficiency of approximately 0.3 kcal mol^{-1}/heavy atom.[32]

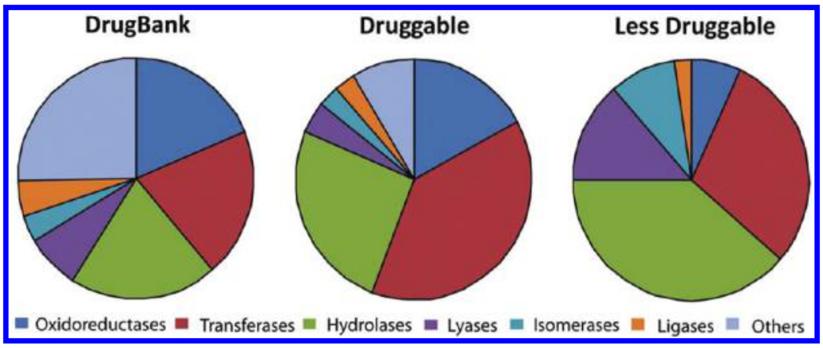


Figure 1. Distribution of the different enzyme classes in the PDB, the druggable and less druggable data set and a subset of DrugBank that only contains one representative per EC number.

Retrospective studies have found that the ligand efficiency is often kept constant during optimization when going from hit to drug.[33] This implies that compounds with ligand efficiencies below 0.3 kcal mol^{-1}/heavy atom are unlikely to be developable into drug molecules (albeit there are exceptions to this rule[34,35]). The PDBbind database was filtered according to these rules to identify potentially less druggable targets. The remaining complexes were cross-checked with Drugbank and ChEMBL to ensure that drug-like ligands were not reported elsewhere. The final set was augmented with the undruggable and difficult targets from the Cheng set.[17] After removing redundant proteins as described above, a set of 44 less druggable protein binding sites was obtained (Table S1).

This NRDLD set contains druggable nuclear hormone receptors, a representative of a G-protein coupled receptor (GPCR) and druggable and less druggable protein binding sites from all six enzyme classes (Figure 1). However, the representation of the enzyme classes in both subsets differs. For instance, oxidoreductases occur more frequently in the druggable subset (17%) than in the less druggable subset (7%). The opposite is true for lyases and isomerases which make up around 14% and 9% of the less druggable subset respectively, but only around 4% and 3% of the druggable ones. Interestingly, enzyme class distributions in the druggable subset do not match the ones found in a nonredundant subset of DrugBank. In addition, nonenzymatic receptors appear to be underrepresented in the druggable subset when compared to established drug targets.

With a total of 115 structures, the NRDLD set is the largest publically accessible nonredundant data set for model building and validation of structure-based druggability methods available to date (Table 1).[16,17,20] All druggable proteins of the Cheng set, some of the Hajduk set, and almost all of the NRDD set are contained in the new data set. However, due to the different definitions and interpretations of druggability for some targets, the classification has changed. For example, β-lactamases were classified by Schmidtke and Barril and Hajduk et al. to be druggable, whereas here they were considered to be less druggable due to the fact that all marketed drugs that target these enzymes are covalent inhibitors and no drug-like inhibitors were reported in HTS campaigns.[6,36] Likewise, nucleoside diphosphate kinase (NDPK) was defined as druggable in the Hajduk data set while here it is included in the less druggable set due to the lack of drug-like inhibitors reported in the literature.

Multivariate Data Analysis and Predictor Evaluation. For each binding site, 16 different descriptors capturing the polarity, size, and the compactness and 40 descriptors capturing the amino acid composition were calculated as described in the Methods section (Table 2). PLS-DA requires normally distributed descriptor values as input variables.[24] Therefore, the Kolmogorov−Smirnov (KS) test was carried out to test the descriptors for normal distribution.[37] For most of the amino acid based descriptors the null hypothesis that the descriptors are normally distributed could be rejected (p-value <0.05), implying that these do not follow a Gaussian distribution. The remaining 22 descriptors that appeared to be normally distributed were used for further analysis (Table 2).

Table 2. Descriptors for Which the Null Hypothesis That They Are Normally Distributed Could Not Be Rejected with a Significance Level of at Least 0.05

descriptor	abbreviation
binding site volume	*vol*
molecular surface area of superligand	*msa*
binding site compactness	*cness*
contact surface area between protein and superligand	*csa*
relative hydrophobic surface area	*hsa_r*
hydrophobic surface area	*hsa_t*
relative hydrogen bond donor surface area	*dsa_r*
hydrogen bond donor surface area	*dsa_t*
relative hydrogen bond acceptor surface area	*asa_r*
hydrogen bond acceptor surface area	*asa_t*
relative polar surface area (dsa_r+asa_r)	*psa_r*
relative occurrence of hydrophobic amino acids	*haa*
relative occurrence of polar amino acids	*paa*
relative occurrence of multi functional amino acids	*maa*
relative occurrence of charged amino acids	*caa*
hydrophobicity index of amino acids	*hiaa*
enrichment of Gly in binding site relative to surface	*fGLY*
enrichment of Val in binding site relative to surface	*fVAL*
enrichment of Leu in binding site relative to surface	*fLEU*
enrichment of Asp in binding site relative to surface	*fASP*
enrichment of Ser in binding site relative to surface	*fSER*
enrichment of Thr in binding site relative to surface	*fTHR*

Principal component analysis (PCA) was performed on the descriptor matrix to investigate the diversity of the binding sites in the NRDLD set. The resulting model had two principal components and an R^2 of 0.61 and a Q^2 of 0.52. Inspection of the score plot (Figure 2a) showed that the binding sites covered the space evenly with no obvious outliers. Less druggable

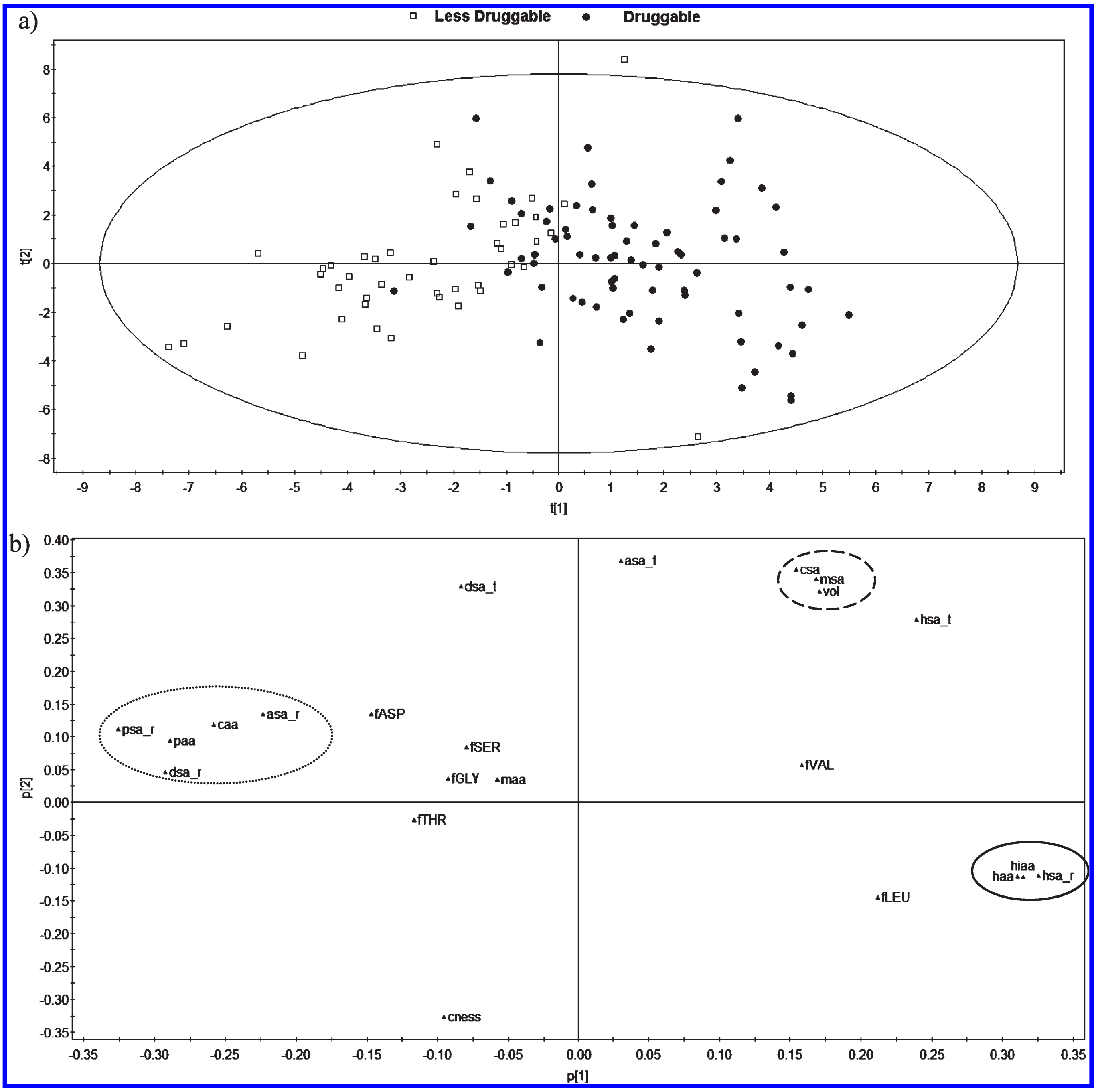


Figure 2. PCA using the druggable and less druggable data set. a) Score plot and b) corresponding loading plot. In both plots the axes represent the first two principal components. a) White squares represent less druggable and black dots druggable binding sites. The black ellipse corresponds to a confidence level of 99% of Hotelling's T2. b) Descriptors describing similar properties are grouped together (dotted circle: polarity descriptors, black circle: hydrophobicity descriptors, dashed circle: size descriptors).

binding sites mostly had negative t1 values and were located on the left-hand side of the plot, whereas druggable binding sites were predominantly found having positive t1 scores. The loading plot (Figure 2b) indicates that the first principal component is mainly derived from differences in active site polarity, whereas the size of the pocket is captured in both the first and the second principal component. According to the PCA, druggable sites show a tendency to be larger and less polar than less druggable proteins which is in agreement with previous findings.[16,20]

Due to the lack of data on experimentally confirmed less druggable targets, previous studies extended this set by including surface clefts identified by pocket detection algorithms which have not necessarily evolved to bind a ligand (here termed "decoy pockets").[15,16,20] To investigate if this approach would be suitable for our purpose, decoy pockets together with their descriptors for all proteins in the NRDLD set were calculated. Subsequently, this data set was merged with the NRDLD set, and a new PCA using the same descriptors as above (Table 2) was carried out. As before, the resulting model had two principal

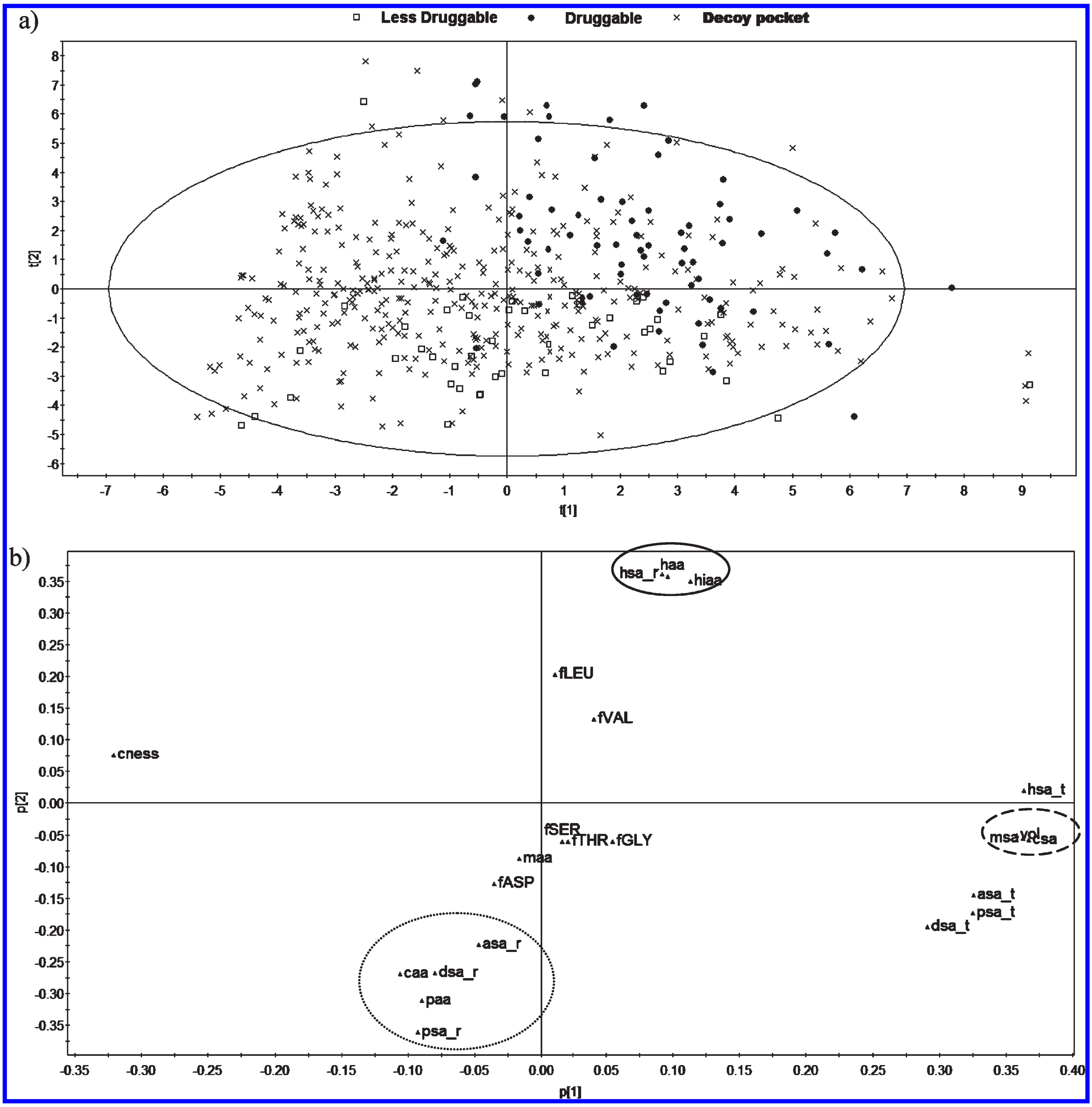


Figure 3. PCA using the druggable and less druggable data set together with decoy pockets. a) Score plot and b) corresponding loading plot. In both plots the axes represent the first two principal components. a) White squares represent less druggable, black dots druggable binding sites and crosses decoy pockets. The black ellipse corresponds to a confidence level of 99% of Hotelling's T2. b) Descriptors describing similar properties are grouped together (dotted circle: polarity descriptors, black circle: hydrophobicity descriptors, dashed circle: size descriptors).

components (Figure 3a) with R^2 and Q^2 values of 0.56 and 0.48, respectively. An overall separation between druggable and less druggable binding sites was still achieved with druggable pockets tending to have positive t2 values, while less druggable pockets tended to have negative t2 values. Again, this was driven by descriptors capturing the polarity of the cavities (Figure 3b). While some of the decoy pockets were located close to druggable or less druggable binding sites in the score plot (Figure 3a), the area with negative t1 values and positive t2 values was almost exclusively occupied by decoy pockets. The latter observation was driven by the compactness, size, and hydrophobicity of the binding sites with decoys pockets located in this area being more spherical and smaller than most druggable binding sites but more hydrophobic than less druggable binding sites (Figure 3b). Taken together, this analysis implies that while some decoy pockets have properties similar to binding sites that have evolved to bind a ligand, others tend to be more compact and smaller than binding sites in general and more hydrophobic than less druggable binding sites.

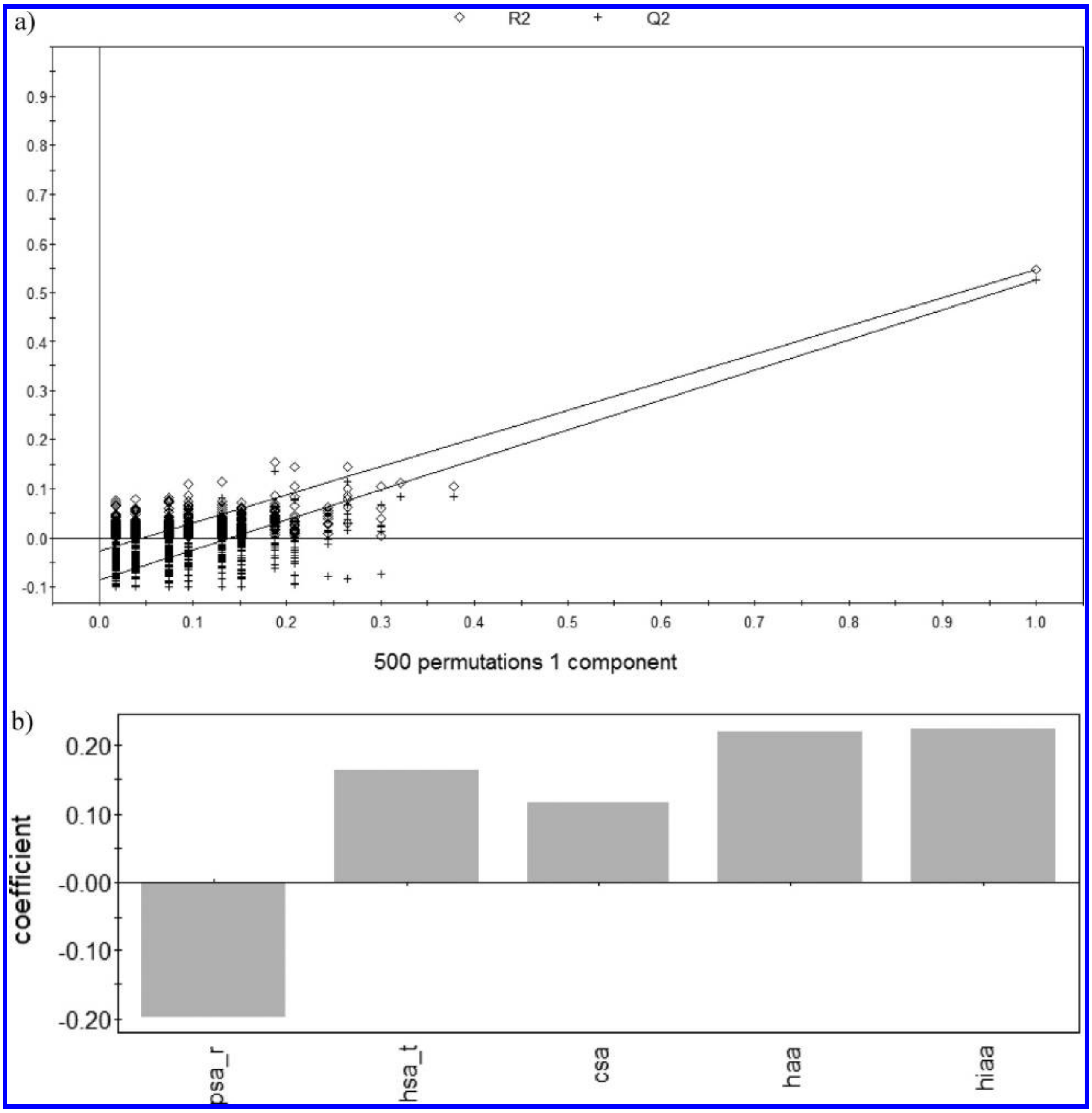


Figure 4. a): Validate plot of the PLS-DA model with 500 Y-permutations. The *y*-axis represents the Q^2 (crosses) and R^2 (diamonds) values of the scrambled and parent PLS-DA models, and the *x*-axis represents the correlation coefficient between the permuted and the original response variable. The Q2 and R2 values of the parent model are found at a correlation coefficient of 1.0. The solid lines represent a regression fit of the data. A crossing of these lines with the *Y*-axis below 0 indicates that the model is not overfitted. b) Coefficient plot for the PLS-DA model. The Y-values represent the influence of each of the descriptors on the druggability score.

The NRDLD set was split into a training and test set to allow training and validation of a PLS-DA model.[24] The model was built using an iterative process starting with the 22 descriptors that appeared to be normally distributed (Table 2). Weak descriptors, which did not show significant contribution to the model, were subsequently omitted, and new models were generated. During this process, hydroxynitrile lyase (HNL) and angiotensin converting enzyme 1 (ACE-1) which were both classified as less druggable were identified as outliers based on the distance to model and removed from the set. The best model (which in the following is called DrugPred) contained the five descriptors *psa_r*, *csa*, *hsa_t*, *haa*, and *hiaa* and had an internal predictivity (Q^2) of 0.54 as judged by cross-validation. However, this value does not provide a measure of the statistical significance of the estimated predictive power. Therefore, multiple PLS-DA models using the same descriptors and components but permuted Y-data were generated. All of these scrambled models lacked predictivity, indicating that the relationships found previously do not stem from spurious correlation (Figure 4a). As expected based on the PCA (Figure 2a), the descriptors *csa*, *hsa_t*, *haa*, and *hiaa* had positive coefficients and therefore a positive influence on the druggability score while *psa_r* had a

Table 3. Accuracy, Recall, and Precision for Druggability Classifications Using Different Prediction Methods and Data Sets

data set (number of nonredundant binding sites)	prediction method	accuracy	recall (druggable/ less druggable)	precision (druggable/ less druggable)	percentage of binding sites in ambiguous region
NRDLD training set (76)	DrugPred[a]	0.91	0.96/0.83	0.90/0.93	0
NRDLD validation set (37)	DrugPred[a]	0.89	0.91/0.85	0.91/0.85	0
NRDLD validation set (37)	Modified DrugPred[a],[b]	0.76	1.00/0.36	0.72/1.00	0
NRDLD training set (76)	DrugPred	0.92	0.95/0.86	0.91/0.93	6.6
NRDLD validation set (37)	DrugPred	0.91	0.91/0.92	0.95/0.86	8.1
DD set (107)	Schmidtke and Barril[20]	0.71[c]	0.88/0.56[c]	0.64/0.83[c]	0
NRDLD validation set (37)	Schmidtke and Barril[20]	0.69	0.68/0.71	0.79/0.59	0
Hajduk set (72)	Hajduk et al[16]	0.58[d]	0.93/0.52[d]	0.66/0.88[d]	19.4
Hajduk set (70)[g]	DrugPred	0.79	0.79/0.80	0.81/0.77	6.0
Cheng set (27)	MAP$_{POD}$[17]	0.96[e]	0.94/1.00[e]	1.00/0.90[e]	
Cheng set (27)	DLID[15]	0.80[f]	0.81/0.87[f]	0.78/0.70[f]	0
Cheng set (26)[g]	DrugPred [a]	0.84	0.94/0.67	0.83/0.80	0
Cheng set (26)[g]	DrugPred	0.91	0.93/0.86	0.93/0.86	19.4

[a] Threshold values for ambiguous predictions were not applied. [b] Derived from a reduced training set. [c] Values calculated based on Figure 3 in ref 20, [d] Figure 5 in ref 16, [e] Figure 2 in ref 17, and [f] Table 1 in ref 15, whereas difficult or prodrug targets were considered to be less druggable. [g] For leghemoglobin and dUTPase from the Hajduk set and ACE-1from the Cheng set, the distance to model was too large to make a reliable prediction. These proteins were therefore removed from the respective sets. In cases where the Cheng set contained multiple structures of a protein the average score was used.

negative contribution (Figure 4b). The three descriptors *hsa_t*, *haa*, and *hiaa* all capture the hydrophobicity of the binding site. However, removing any of them resulted in worse models (data not shown). It seemed clear from the PCA plots (Figure 3) that decoy pockets should not be included in the data set when deriving a PLS-DA model. This was confirmed when models were built using a data set consisting of the training set binding sites and decoy cavities taken together. The internal predictivity of the best model derived in this way was very low (Q^2 = 0.15, R^2 = 0.16). On using this model to make predictions for the test set, the highest score obtained was ~0.4, i.e., all the pockets scored in the less druggable zone. Further, we investigated the influence of the size of the data on the outcome of the predictions. Binding sites from the training set were randomly removed to obtain a set that resembled the one used by Schmidtke and Barril to train their druggability predictor (35 druggable and 8 less druggable sites).[20] New PLS-DA models using this reduced set were subsequently built. The best model performed comparably to DrugPred on the training set (data not shown) but had worse accuracy, precision, and recall for the validation set (Table 3).

Overall, the DrugPred model was able to separate druggable and less druggable binding sites (Figure 5). Most druggable binding sites were assigned a score close to 1, while most less druggable ones obtained a score close to 0. The model has a good predictive power for both internal and external predictions with accuracy around 0.9 for both, training and validation sets, recall ranging from 0.83 to 0.96 and precision from 0.85 to 0.93 (Table 3). However, an ambiguous region was observed in the middle of the scale, where a separation of druggable and less druggable sites was not possible (Figure 5). Therefore, threshold values were introduced to separate the ambiguous region from the druggable and less druggable regions based on the score distributions of the true positive and negative predictions. Discounting the binding sites in the ambiguous region increased accuracy (0.91 and 0.92, respectively), recall (0.86−0.95), and precision (0.86−0.95).

Next, the robustness of DrugPred against conformational changes in the binding site was evaluated. For that purpose, a data set of protein binding sites for which different conformations were found when cocrystallized with different ligands was used (Table S2). This data set was a modification of a set originally developed to validate the flexible docking program FlexE.[38] The original data set contained ten different proteins with various degrees of flexibility in the binding sites. In addition, some examples of the proteins in the FlexE set contained mutations. Since changing the physicochemical properties of the pockets can lead to a change of druggability prediction, the mutated proteins were removed from the set and, if possible, replaced with structures of the same protein free of amino acid replacements. The DrugPred predictions for all structures in the modified set were calculated, and the average scores together with the minimum and maximum scores for each protein were plotted (Figure 6). For seven out of the ten proteins in the data set all binding site conformations were predicted to belong to the same category. For the remaining proteins, the average score predicted the binding site to be less druggable, but at least one representative of these proteins belonged to a different category. The magnitude of the change of the DrugPred score was not necessarily dependent on the magnitude of the structural changes in the binding sites but on their consequences (Table S2). For instance, the pairs of structures 1dyi and 1ra3 (dihydrofolate reductase, DHFR), 1ah0 and 1ah3 (aldose reductase, AR), and 1ahb and 1mom (alpha-momorcharin, AM) all had rmsd values for C-alpha atoms in the binding sites between 0.212 and 0.401 Å and one or two side chain movements. The binding site changes for the DHFR pair had only little consequence for the size and polarity of the binding pocket (Figure 7a and b). Accordingly, the DrugPred score was similar for both binding sites (1.02 vs 0.98). For the other two pairs, the conformational changes altered the size of the binding pockets drastically (Figure 7c-f). For AR, both, the small and large binding site, were scored as druggable (0.78 vs 0.77). In contrast, for AM, only the larger pocket was predicted to have the required

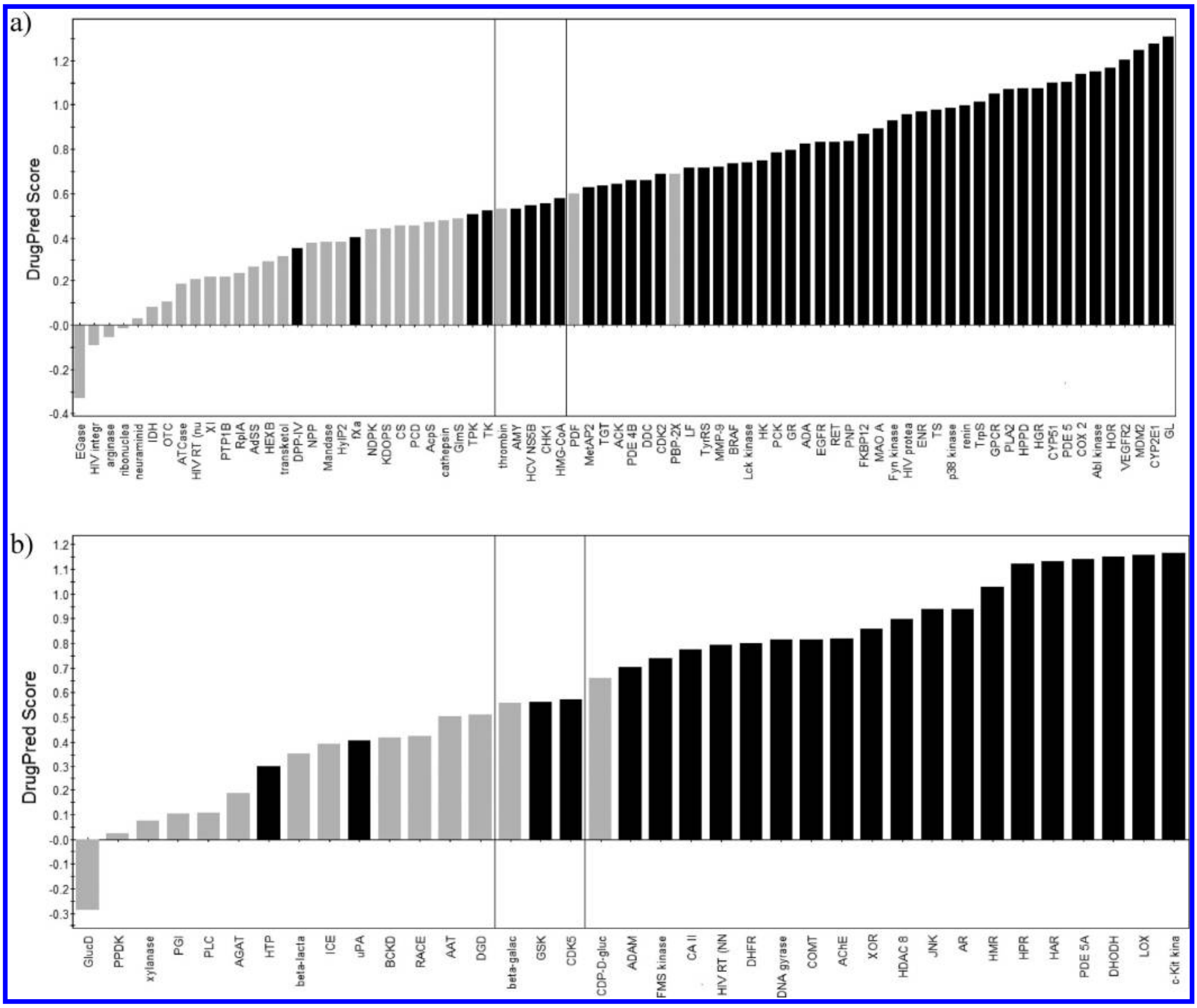


Figure 5. a) All structures in the training (a) and validation set (b) ordered according to their DrugPred score. The borders that separate the ambiguous from the less druggable and druggable region are indicated as vertical lines. Druggable binding sites are marked in black and less druggable ones in gray.

size and polarity to bind orally available ligands with high affinity (0.61 vs 0.42).

Finally, the effect of the ligand used to seed the descriptor calculations on the DrugPred score was tested. The smallest and largest ligands for any given protein of the modified FlexE set (Table S 2) were extracted and merged with two copies of the same structure containing only protein and, where applicable, the cofactor in order to nullify any side chain movements that might affect the score. Descriptors were then calculated based on the generated complex structures and used as input for DrugPred. Albeit the ligand size varied by up to 29 heavy atoms for a given protein, in all investigated examples the same prediction was obtained.

Comparison of DrugPred with Other Druggability Prediction Methods. The performance of DrugPred was compared to that of other previously published druggability prediction methods. For this purpose, the predictive power of druggability methods that not only assign a continuous score but also classify binding sites into druggable and less druggable ones was calculated based on the validation data presented in the respective publications (Table 3).[15−17,20] In addition, the structures in the Hajduk and Cheng sets were scored by DrugPred, and the structures in the NRDLD validation set were scored using the method developed by Schmidtke and Barril.[20] The accuracy achieved by DrugPred is between 0.04 and 0.21 units higher than the accuracy achieved by DLID or either of the methods developed by Schmidtke and Barril or Hajduk et al., depending on which set is used for comparison and if the threshold values for the ambiguous region of DrugPred are taken into account or not. The higher accuracy is reflected in general higher recall and precision values. The only data available for MAP_{POD} is the performance on the relatively small Cheng set that was used to derive the method.[17] Using this set, MAP_{POD} achieves a slightly higher accuracy than DrugPred (0.96 vs 0.91).

DISCUSSION

A new structure-based druggability predictor (DrugPred) was developed. DrugPred is robust against conformational changes in the binding sites (Figure 6), independent of the ligand bound in

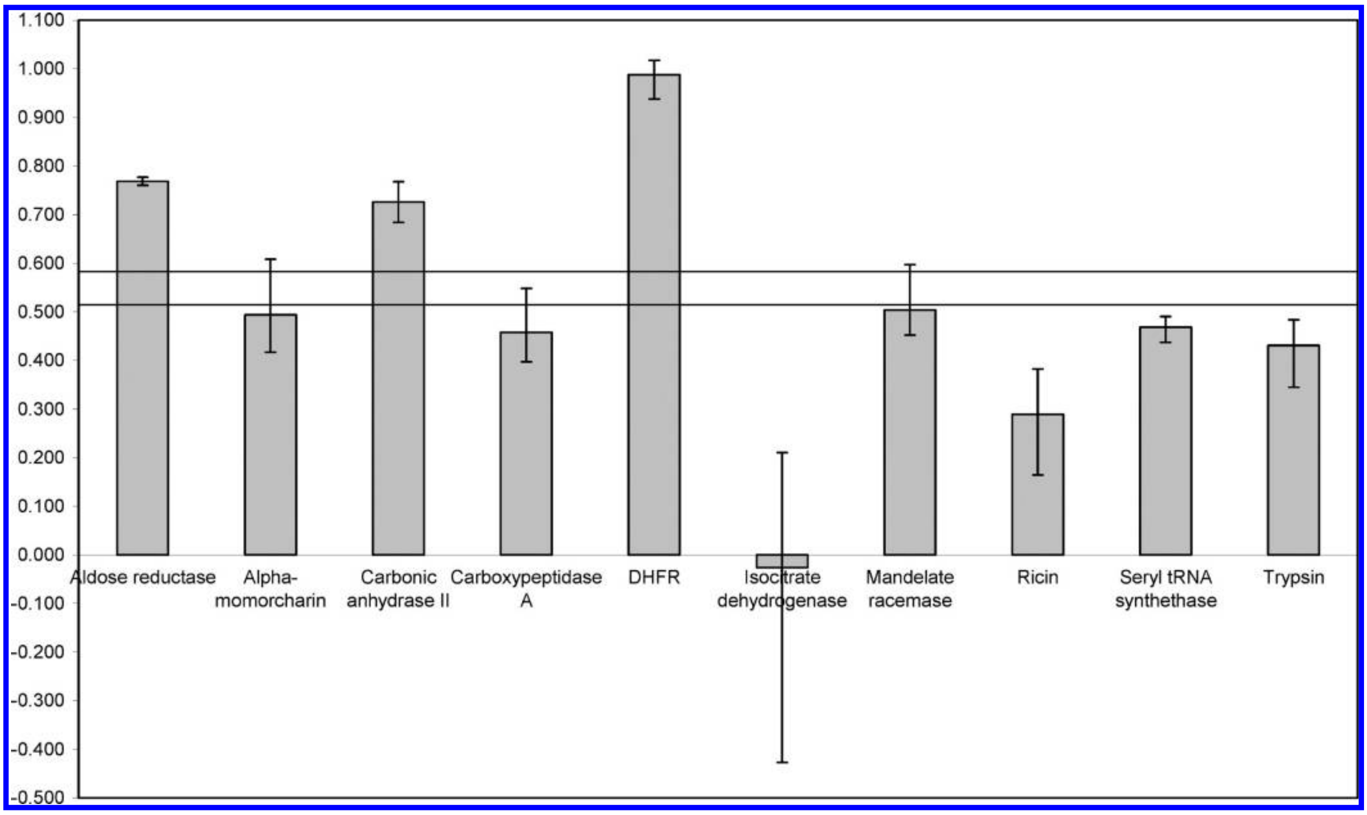


Figure 6. Average scores of the structures in the modified FlexE set with error bars representing the minimum and maximum score. The borders between druggable, ambiguous, and less druggable regions are marked with horizontal lines.

the input structure, and outperforms all but one previously published method (Table 3). Since the descriptors that went into the druggability model are similar to the ones used previously, we believe that the carefully chosen data set of druggable and less druggable binding sites together with the decision to not include decoy pockets was key to success. Despite the good performance, DrugPred is not free of wrong predictions (Figure 5). To further improve the method, it is important to distinguish between wrong classifications and real miss-predictions. We expand on all of these points in the following.

We compiled a new comprehensive data set of druggable and less druggable binding sites to facilitate development and validation of structure-based druggability methods (Table S1). This data set compromises 115 binding sites and is the largest nonredundant data set available in the public domain (Table 1). The fact that all enzyme classes are represented in both the druggable and less druggable subset makes it well suited for model building (Figure 1). Interestingly, the distribution of enzyme classes between both subsets differs and does not match the distribution found among established drug targets. The latter is likely due to the availability of determined protein structures of drug targets. It was estimated that marketed drugs have about 200−400 different primary targets.[7,25,26] Not all of them have been crystallized to date, and, to avoid redundancy, only 71 of the available structures were included in the final set. Further, enzymes only account for a fraction of known drug targets. Other important protein families are NHRs which constitute about 3% of known targets, ion channels (4%), and the large group of rhodopsin-like GPCRs (25%).[7] The ligand-binding domains of NHRs are amenable for crystallization, and the structures of a number of important drug targets have been determined.[39] As a consequence, they correspond to about 7% of binding sites in the druggable subset presented here. In contrast, the membrane bound ion channels and GPCRs are rather difficult to crystallize, and structures of drug targets are only just emerging.[40,41] While the binding site of the $\beta 2$ adrenergic receptor was included in the druggable set, the lack of ion channel structures in complex with drug-like ligands made it impossible to consider them. Besides waiting for more crystal structures to become available, another option to enhance the data set in the future is to use the available GPCR structures as templates for homology modeling and to include the models into the set.

The druggability predictor DrugPred was developed based on the compiled data set. The method shows high accuracy (>0.89), recall (>0.83), and precision (>0.85) when discriminating druggable from less druggable binding sites (Table 3). Among the 56 calculated binding site descriptors (Table 2) those capturing the size (*csa*), polarity (*psa_r*), and hydrophobicity (*hsa_t*, *haa*, and *hiaa*) were found to be most important for the model (Figure 4b). While *haa* and *hiaa* are only dependent on the amino acid composition of the binding site, *hsa_t* also includes contributions from cofactors. This might explain why removing any of these three descriptors from DrugPred leads to a worse model. The final descriptors capture similar properties as were found to be crucial in previous work.[15−18,20] This is less surprising considering that drug-like ligands are often described by the Rule of Five which also captures polarity and size of the compounds.[11] Accordingly, binding sites accommodating this type of compounds must have complementary properties. However, somewhat unexpected the shape descriptor binding site compactness (*cness*) turned out not to be significant to separate

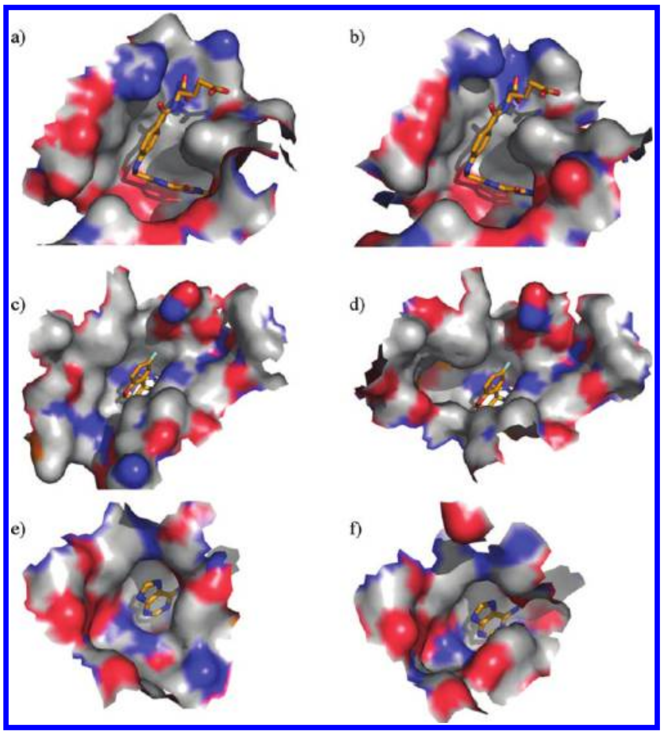


Figure 7. Solvent accessible surface of the binding sites for selected examples from the modified FlexE set. Surface patches originating from nitrogen atoms are colored in blue, from oxygen atoms in red, and from carbon atoms in gray. For clarity, the ligands used to determine the binding site location are also shown. a and b): DHFR binding site of 1dyi and 1ra3 (b). The conformational change has little influence on shape and polarity of the pocket. c and d) AR binding site of 1ah0 (c) and 1ah3 (d). The conformational change alters the size of the pocket. Both binding sites are predicted to be druggable. e and f) AM binding site of 1ahb (e) and 1mom(f). Only the larger binding site (f) has the required size and polarity to score as druggable. (Figures made using PyMOL (Schrödinger).).

druggable from less druggable binding sites. Rule of Five compliant drugs were found to rarely be sphere shaped.[42] Therefore, we anticipated that a descriptor which encodes how closely a binding site resembles a sphere would be important for a structure-based druggability prediction model. Whether the descriptor was not sufficient for describing the shape of the binding site and other shape descriptors such as the ones based on principal moments[16,43] are more appropriate or if the shape in general is not a useful feature to distinguish druggable from less druggable binding sites is a matter for future investigations. Further, in order to not confuse the pocket identification problems with druggability predictions, DrugPred was deliberately designed to not detect binding sites automatically but to rely on user input. The most straightforward way to indicate the binding site location is the presence of a bound ligand, but output from pocket identification programs such as FPOCKET[44] can also be used. Therefore, it is crucial that the predictions are independent of the choice of the ligand as demonstrated using examples from the modified FlexE set (Table S2). In addition, the method is robust against conformational changes in the binding sites (Figure 6). In most of the investigated proteins, different binding site conformations did not lead to different binding site classifications. That for some proteins a different classification was obtained is not necessarily a weakness of the method. Conformational changes that enlarge a binding site and/or change its physicochemical properties (Figure 7e and f) might well be a prerequisite for binding a drug-like molecule. It is therefore advisable, to score multiple binding site conformations for target assessment, especially if some of the pockets obtain a score close to the ambiguous region.

DrugPred performs better than most previously published structure-based druggability prediction methods (Table 3). The only method with a better predictive power is $\mathrm{MAP_{POD}}$[17] when applied to the same test set. However, it has to be considered that the $\mathrm{MAP_{POD}}$ score limit dividing druggable from less druggable or difficult targets was manually adjusted to achieve ideal performance on this data set. For a fair comparison, $\mathrm{MAP_{POD}}$ predictions on an independent data set would be needed. Compared to other regression based druggability methods,[15,16,20] DrugPred performs better in terms of accuracy, precision, and recall despite using similar descriptors (Table 3). In all four models the hydrophobicity/polarity of the binding site together with the volume were found to be important to distinguish druggable from less druggable binding sites (Figure 4b). An important difference in deriving the druggability prediction methods was the size of the training set and its composition. First, the data set described here is the largest nonredundant data set that was used to train a druggability predictor. When a subset with the same size and ratio of druggable to less druggable binding sites as published previously[20] was used for model building the predictions for the validation set got worse (Table 3). In particular affected were the results for less druggable binding sites. While all binding sites that were predicted to be less druggable belonged indeed to this category (precision = 1.00) a large number of less druggable binding sites was wrongly classified to be druggable (recall = 0.36). This is probably due to the low number and consequently limited diversity of less druggable binding sites in the reduced training set. Second, the NRDLD set is free of surface clefts which have not evolved to bind a ligand. For training their predictors, Hajduk et al., Schmidtke and Barril, and Sheridan et al. enlarged their less druggable binding site set with such decoy pockets, whereas we deliberately decided not to do so. A PCA on a data set containing decoy pockets as well as druggable and less druggable binding sites shows that some decoy pockets are close in property space to active sites while others have distinct characteristics (Figure 3). This is in agreement with previous observations that surface clefts have different properties than ligand binding sites.[45] In fact, this is the principle on which most pocket detection algorithms are based. By including such non ligand binding clefts into the data set, especially when outnumbering ‘real’ less druggable binding sites as done previously,[15,16,20] the model will actually be trained for the wrong purpose. This effect can be seen when including decoy pockets as less druggable binding sites into the training set. As expected, in this scenario the predictions for druggable binding sites became worse. In summary, the larger size of the training set together with the exclusion of decoy pockets might explain the superior performance of DrugPred compared to the other methods.

Despite its good performance DrugPred is not free of errors (Figure 5). The less druggable subset had to be derived indirectly via the absence of published drug-like ligands due to the lack of data on confirmed less druggable drug targets. As a result, we expected that there might be a number of false positives in this subset. It is therefore rather surprising, that fewer less druggable protein binding sites were predicted to be druggable than *vice versa*. In the training set, the binding sites of peptide deformylase

(PDF) and penicillin binding protein-2X (PBP-2X) were wrongly classified as druggable, whereas dipeptidyl peptidase-IV (DPP-IV), factor Xa (fXa), thiamine pyrophosphokinase (TPK), and thymidine kinase (TK) were misclassified as less druggable (Figure 5a). In the test set, CDP-D-glucose synthase (CDP-D-gluc) was incorrectly classified as druggable and urokinase plasminogen activator (uPA) and human thymidine phosphorylase (HTP) as less druggable (Figure 5b). The enzymes uPA and HTP were included in the druggable subset because they are part of the Astex diverse set and TPK because it is part of the Astex diverse set and DrugBank.[28,29] Neither of the proteins is a clinically used drug target. The only approved drug associated in Drugbank with TPK is thiamine which is the substrate of the enzyme. No inhibitors of TPK are deposited in ChEMBL.[31] Based on these data, it remains unclear if TPK is actually druggable, and it is advisible that for further studies this target is removed from the druggable set. For uPA and HTP inhibitors are deposited in ChEMBL. Most inhibitors of uPA possess a highly basic guanidine or amidine group making it unlikely that they are orally available. Likewise, the most potent inhibitors of HTP are either charged or require administration as prodrugs.[46,47] Both binding sites appear therefore to be misclassified and should rather be added to the less druggable subset for future studies. In contrast, for the wrongly predicted enzymes DDP-IV, fXa and TK orally available drugs exist.[23,48] TK is the target for antiviral compounds such as acyclovir and ganciclovir.[49] Accordingly, this enzyme is druggable and was wrongly predicted by DrugPred. The four clinically used DPP-IV inhibitors are rule of 5 compliant small molecules.[48] However, three of them are reversible covalent inhibitors and therefore do not comply with the druggability definition used in this work. The structure included in the druggable subset contains a side chain movement of Arg358 compared to the complex cocrystallized with sitagliptin, the noncovalent drug (PDB code 1X70). As a consequence, the binding site is enlarged (not shown). When this structure is used for druggability prediction, the DrugPred score changes from less druggable to ambiguous. This observation highlights again that even if DrugPred is rather robust toward conformational changes in the binding site, it is advisible to use more than one binding site conformation for target assessment if available. The development of orally available fXa inhibitors was a long journey. A major challenge was to find an arginine mimic that did not contain a highly basic center and thus contributed to poor oral bioavailability.[23] Not only DrugPred but also DLID and the druggability method developed by Schmidtke and Barril fail to predict this target correctly.[15,20] This reemphasizes the fact that the druggability concept is relative, and with a lot of effort and resources also for difficult targets an orally available compound could potentially be developed.

CONCLUSION

A large data set for assessing protein druggability was compiled. This set extends the scope of the previously published sets[16,17,20] by including more nonredundant structures and a larger variety of less druggable protein binding sites. Using this set, we derived and validated a new druggability model that successfully classified binding sites as being druggable or less druggable. We are currently working on making this method available via a Web server. Further, encouraged by the robustness of DrugPred against conformational changes in the binding sites, we are evaluating its performance on homology models which usually have some degree of uncertainties, especially for side chain conformations. Of course, druggability assessment of binding sites is only one criteria besides others when deciding on which projects to work on and how many resources to allocate.[9] Its particular value comes into play when the task is to prioritize the most promising targets among a long list of potential targets, for example when analyzing a whole genome.[50] Therefore efforts are under way to make the method robust for high-throughput predictions to facilitate this task.

METHODS

Calculation of Ligand Properties. Ligand efficiency[32] was calculated as the ratio of the binding energy to the number of heavy atoms. If no K_d to calculate the binding energy was stored in the PDBbind database[30] (released 2006) the K_i or IC_{50} values were used instead as approximations. ClogP was calculated using the Sybyl Programing Language (SPL) expression generator *% clogp* as implemented in SYBYL −X 1.0 (Tripos, St. Louis, Missouri, USA). Molecular weight and number of hydrogen bond acceptors and donors were calculated using the OEChem python library (version 1.7.2) by OpenEye (Santa Fe, NM, USA).

Compilation of the NRDLD and Modified FlexE Data Sets. The druggable targets in the NRDRD set (Table S1) were retrieved by mining the relevant literature[25−27] and DrugBank[28] and augmented with the structures from the Astex diverse set[29] and the druggable proteins from the Cheng set.[17] The less druggable targets were obtained by filtering all protein−ligand complexes with a resolution better than 2.6 Å in PDBbind (released 2006) for the following criteria: At least one of the ligands cocrystallized with the same protein did fulfill the rule of five, had a clogP ≥ −2 and a ligand efficiency[32] ≥ 0.3 kcal mol^{-1}/heavy atom. All proteins matching these criteria were considered to be potentially druggable and rejected. The remaining complexes were cross-checked with Drugbank and ChEMBL[31] to ensure that no drug-like ligands were reported elsewhere. The final set was augmented with the undruggable and difficult targets from the Cheng set.[17] If more than one structure of the same protein was deposited in the PDB, the structure with the highest resolution and, in the case for druggable proteins, containing preferably a drug-like ligand, was included in the set. Subsequently, for both the druggable and less druggable subset of the NRDLD a sequence alignment was performed for enzymes having the same EC number, and structures with a sequence identity higher than 60% to other members in the same group were filtered out in order to avoid redundancy.

For the modified FlexE set (Table S2), all structures in the original FlexE set[38] that contained mutations in the binding site were removed. If possible, they were subsequently replaced with structures of the same protein free of amino acid changes in the binding sites retrieved via Relibase+.[51] The same ligand was used for all structures of identical proteins as matching spheres to generate the superligand (see below). Structures in which this sphere set resulted in a clash with the protein were rejected from the set. If cofactor and/or catalytic metal ions were missing from one structure, the coordinates were taken from another structure of the same protein.

Binding Site Boundaries. To establish the boundaries of a binding pocket, all approved drug molecules taken from DrugBank[28] for which a dockable database format could be generated (995 out of 1024) were docked into the pocket using

DOCK 3.5.54.[52] The atoms of the ligands bound in the crystal structures were used as matching points for docking, and all water molecules and ligands except for cofactors and catalytically important ions were removed. Default settings were used with ligand and receptor bins set to 0.5 Å, overlap bins set to 0.4 Å, and the distance tolerance for matching ligand atoms to receptor matching sites ranged from 1.1 to 1.2 Å. Since the aim of docking was solely to obtain information about the shape and the volume of the binding sites, all protein atoms were set to carbon atoms and assigned a partial charge of 0. Subsequently, all compounds for which a docking pose was obtained and for which the ratio of VDW score to number of heavy atoms was ≤ -1.2 were merged into a superligand, serving as a negative print of the binding site. Based on this print binding site descriptors were calculated as follows.

Binding Site Descriptor Calculation. All water molecules and ligands except for cofactors and catalytically important ions were removed before descriptor calculation.

Volume, Size, and Polarity. The volume (*vol*) and surface area (*msa*) of the binding site were calculated directly from the superligand using the SPL expression generators *%volume* and *%surface*. To calculate the contact surface area (*csa*) of the superligand and binding site, the space behind binding site atoms in contact with bulk solvent was blocked with dummy atoms. Next, the surface area of all binding site atoms within 2.1 Å of the superligand was calculated using SYBYL-X, and its size was determined using the SPL expression generator *%surface*. Using the MOLCAD module in SYBYL-X hydrogen-bond donor and acceptor properties of the binding sites were mapped onto *csa*, and their size were determined delivering the descriptors total hydrogen-bond donor (*dsa_t*) and acceptor surface area (*asa_t*). Based on these descriptors, the total hydrophobic surface area (*hsa_t*) was defined as the difference between *csa* and the sum of *dsa_t* and *asa_t*. In addition, the relative size of the hydrogen-bond donor (*dsa_r*), acceptor (*asa_r*), polar (*psa_r*), and hydrophobic surface area (*hsa_r*) were obtained by dividing each of the descriptors by *csa*.

Binding Site Compactness. The compactness of a binding site was defined as the ratio between the minimum possible surface area (A_{sph}) for the volume (*vol*) of the superligand and the actual molecular surface area (*msa*). The minimum surface area to cover a given volume is the one of a sphere. Therefore, eq 1 was used to obtain the radius of a sphere (r_{Vsph}) with the same volume (*vol*) as the superligand. In the second step, the area of a sphere A_{sph} with the radius (r_{Vsph}) was calculated using eq 2

$$r_{V_{sph}} = \sqrt[3]{\frac{vol}{\frac{4}{3}\pi}} \tag{1}$$

$$A_{sph} = 4\pi(r_{V_{sph}})^2 = 4\pi\left(\sqrt[3]{\frac{vol}{\frac{4}{3}\pi}}\right)^2 \tag{2}$$

By substituting A_{sph} into eq 3 the compactness descriptor (*cness*) was obtained. According to this definition, the closer *cness* is to 1, the more compact (spherical) is the binding site

$$cness = \frac{A_{sph}}{msa} = \frac{4\pi\left(\sqrt[3]{\frac{vol}{\frac{4}{3}\pi}}\right)^2}{msa} \tag{3}$$

Amino Acid Descriptors. All amino acids having their side chains oriented toward the ligand and being not further away than 4 Å from a superligand atom were extracted. The amino acids were grouped by their overall physicochemical properties, i.e. Ala, Gly, Val, Ile, Leu, Met, Phe, and Pro were considered to be apolar; Thr, Lys, Arg, Glu, Asp, Gln, Asn, and Ser to be polar; Lys, Arg, His, Asp, and Glu to be charged; and Trp, Tyr, His, and Cys to be multifunctional. The fraction of each group of amino acids with respect to the total number of amino acids in the binding site was calculated and used as the descriptors hydrophobic amino acids (*haa*), polar amino acids (*paa*), multifunctional amino acids (*maa*), and charged amino acids (*caa*). The hydrophobicity of the binding site was captured by the descriptor *hiaa* representing the sum of the hydrophobicity indices of the amino acids lining the binding site as defined by Kyte and Doolittle.[53] Finally, the relative enrichment of each amino acid in the binding site as compared to its average occurrence on a protein surface was calculated as described by Soga et al.[19]

Decoy Pocket Detection. The program FPOCKET[44] was used to identify the ten best scoring pockets on each protein structure in the NRDLD set using the default parameters (minimum and maximum α-sphere radii of 3 Å and 6 Å, respectively; pockets defined as a minimum of 35 α-spheres, which were clustered together if within 1.73 Å of each other; α-sphere clusters whose center of mass are within 4.5 Å of each other were also clustered together). To generate spheres for docking, the van der Waal's energy of each α-sphere was calculated. The most buried α-sphere was identified as the one with the least energy. The distance between the most buried α-sphere and the one farthest from it ($\mathrm{MaxD_0}$) was then identified. All α-spheres at a distance of more than 25% of $\mathrm{MaxD_0}$ from the most buried α-sphere were eliminated. If the remaining spheres were closer than 3 Å to the ligand bound in the crystal structure, the pocket was rejected. This procedure resulted in 392 decoy pockets.

Test for Normal Distribution. The Kolmogorov−Smirnov (KS) test was performed to check if a descriptor was normally distributed by using the program SPSS (Superior Performing Software System, Chicago, IL, USA).[37] The null hypothesis (H_0) was defined as follows: if a descriptor sample $F_{descr}(x)$ has the same distribution as a reference sample $F_{norm}(x)$ then it is normally distributed.

Multivariate Data Analysis. Multivariate data analysis was preformed with the program SIMCA-P+ 12.0.1 (Umetrics, Sweden).

Principal Component Analysis (PCA). The descriptor matrix was normalized to unit variance before carrying out PCA via the PCA-X option from the Sicma-P+ package using standard settings. The number of components extracted was based on cross-validation and eigenvalue analysis.

Partial Least-Squares Projection to Latent Structures Discriminant Analysis (PLS-DA). All calculated descriptors were normalized to unit variance. In addition, the descriptors *paa*, *dsa_t*, *asa_t*, and *dsa_r* and all enriched amino acid descriptors were skewed and therefore corrected by logarithmic or exponential transformation. If not otherwise stated, all proteins in the compiled data set (Table S1) were ordered by EC numbers and every first and second protein added to the training set and every third to the test set. Receptors were randomly split between training and test set. This resulted in 76 binding sites in the training set (48 of them druggable and 28 less druggable) and 37 in validation set (23 druggable and 14 less druggable). Subsequently, the druggable binding sites were assigned to 1 and the less druggable ones to 0, and a predictive model was built using an iterative process: Weak descriptors, which did not show significant contribution to the model, were identified by consulting the variable importance plot, the loading plot, and the internal predictive power and removed. Outliers were identified based on the distance to the model in the X space (DmodX) and removed. New models were subsequently generated using the reduced number of descriptors. The final model (Figure 4b) was described by the following equation

$$\begin{aligned}\mathrm{DrugPred} = & -0.2{*}psa_r + 0.16{*}hsa_t + 0.11{*}csa \\ & + 0.22{*}apaa + 0.22{*}hiaa + 1.3 \end{aligned} \tag{4}$$

Threshold values for the ambiguous region were calculated as follows: The predicted druggability scores of the druggable and less druggable structures in the training set were used to define a one-sided 90% cutoff value for each class. Assuming an approximate normal distribution of the scores, the thresholds were set at 1.28 times the respective standard deviations added or subtracted from the mean of either class, respectively. This yielded an ambiguous region between 0.51 and 0.59.

Characterization of the Predictive Power. Accuracy (eq 5), precision (eq 6), and recall (eq 7) were calculated to assess the predictive power of the derived druggability prediction models whereas true positives (*tp*) is the number of correctly classified and false positives (*fp*) the number of wrongly classified targets of the predicted class, true negatives (*tn*) is the number of targets correctly classified and false positives (*fp*) the number of targets wrongly classified to not belong to the class

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{5}$$

$$precision = \frac{tp}{tp + fp} \tag{6}$$

$$recall = \frac{tp}{tp + fn} \tag{7}$$

Scripts to calculate the descriptors are available from the authors upon request.

ASSOCIATED CONTENT

Supporting Information. List of the PDB codes of the proteins in the NRDLD set. Table with the modified FlexE set together with rmsd values for binding site residues, number of side chain movements, and DrugPred score. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone +44 1302 386230. E-mail: r.brenk@dundee.ac.uk.

ACKNOWLEDGMENT

A.K. was supported through a Collaborative Awards in Science and Engineering (CASE) studentship, jointly funded by AstraZenca and the Biotechnology and Biological Sciences Research Council (BBSRC). Further funding came from the EC Seventh Framework Programme (FP7/2007-2013, grant agreement n° 223461) and the Welcome Trust (WT083481). We thank Openeye for free software licenses and Jonathan Downward and Tom Walsh for administration of the workstations and high performance computing cluster.

REFERENCES

(1) Chanda, S. K.; Caldwell, J. S. Fulfilling the promise: drug discovery in the post-genomic era. *Drug Discovery Today* **2003**, *8*, 168–174.

(2) Arcus, V. L.; Lott, J. S.; Johnston, J. M.; Baker, E. N. The potential impact of structural genomics on tuberculosis drug discovery. *Drug Discovery Today* **2006**, *11*, 28–34.

(3) Buchanan, S. G.; Sauder, J. M.; Harris, T. The promise of structural genomics in the discovery of new antimicrobial agents. *Curr. Pharm. Des.* **2002**, *8*, 1173–1188.

(4) Van Voorhis, W. C.; Hol, W. G.; Myler, P. J.; Stewart, L. J. The role of medical structural genomics in discovering new drugs for infectious diseases. *PLoS Comput. Biol.* **2009**, *5*, e1000530.

(5) Weigelt, J.; McBroom-Cerajewski, L. D.; Schapira, M.; Zhao, Y.; Arrowsmith, C. H. Structural genomics and drug discovery: all in the family. *Curr. Opin. Chem. Biol.* **2008**, *12*, 32–39.

(6) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 29–40.

(7) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.

(8) Hajduk, P. J.; Huth, J. R.; Tse, C. Predicting protein druggability. *Drug Discovery Today* **2005**, *10*, 1675–1682.

(9) Wyatt, P. G.; Gilbert, I. H.; Read, K. D.; Fairlamb, A. H. Target validation: linking target and chemical properties to desired product prof. *Curr. Top. Med. Chem.* **2011**, *11*, 1275–1283.

(10) Chen, H.; Yang, Y.; Engkvist, O. Molecular topology analysis of the differences between drugs, clinical candidate compounds, and bioactive molecules. *J. Chem. Inf. Model.* **2010**, *50*, 2141–2150.

(11) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(12) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(13) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250–1256.

(14) Varma, M. V.; Obach, R. S.; Rotter, C.; Miller, H. R.; Chang, G.; Steyn, S. J.; El-Kattan, A.; Troutman, M. D. Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. *J. Med. Chem.* **2010**, *53*, 1098–1108.

(15) Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; 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.

(16) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518–2525.

(17) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(18) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.

(19) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **2007**, *47*, 400–406.

(20) 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.

(21) Potashman, M. H.; Duggan, M. E. Covalent modifiers: an orthogonal approach to drug design. *J. Med. Chem.* **2009**, *52*, 1231–1246.

(22) Hsieh, P. W.; Hung, C. F.; Fang, J. Y. Current prodrug design for drug discovery. *Curr. Pharm. Des.* **2009**, *15*, 2236–2250.

(23) Perzborn, E.; Roehrig, S.; Straub, A.; Kubitza, D.; Misselwitz, F. The discovery and development of rivaroxaban, an oral, direct factor Xa inhibitor. *Nat. Rev. Drug Discovery* **2011**, *10*, 61–75.

(24) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(25) Imming, P.; Sinning, C.; Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discovery* **2006**, *5*, 821–834.

(26) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

(27) Zheng, C. J.; Han, L. Y.; Yap, C. W.; Ji, Z. L.; Cao, Z. W.; Chen, Y. Z. Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* **2006**, *58*, 259–279.

(28) Wishart, D. S.; Knox, C.; Guo, A. C.; 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*, D668–672.

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

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

(31) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, [Epub ahead of print].

(32) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.

(33) Hajduk, P. J. Fragment-based drug design: how big is too big? *J. Med. Chem.* **2006**, *49*, 6972–6976.

(34) Perola, E. An Analysis of the Binding Efficiencies of Drugs and Their Leads in Successful Drug Discovery Programs. *J. Med. Chem.* **2010**, *53*, 2986–2997.

(35) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand binding efficiency: trends, physical basis, and implications. *J. Med. Chem.* **2008**, *51*, 2432–2438.

(36) Bebrone, C.; Lassaux, P.; Vercheval, L.; Sohier, J. S.; Jehaes, A.; Sauvage, E.; Galleni, M. Current challenges in antimicrobial chemotherapy: focus on ss-lactamase inhibition. *Drugs* **2010**, *70*, 651–679.

(37) Drew, J. H.; Glen, A. G.; Leemis, L. M. Computing the cumulative distribution function of the Kolmogorov-Smirnov statistic. *Comput. Stat. Data Anal.* **2000**, *34*, 1–15.

(38) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations *J. Mol. Biol.* **2001**, *308*, 377–395.

(39) Huang, P.; Chandra, V.; Rastinejad, F. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. *Annu. Rev. Physiol.* **2010**, *72*, 247–272.

(40) Cherezov, V.; Abola, E.; Stevens, R. C. Recent progress in the structure determination of GPCRs, a membrane protein family with high potential as pharmaceutical targets. *Methods Mol. Biol.* **2010**, *654*, 141–168.

(41) Corringer, P. J.; Baaden, M.; Bocquet, N.; Delarue, M.; Dufresne, V.; Nury, H.; Prevost, M.; Van Renterghem, C. Atomic structure and dynamics of pentameric ligand-gated ion channels: new insight from bacterial homologues. *J. Physiol.* **2010**, *588*, 565–572.

(42) Akritopoulou-Zanze, I.; Metz, J. T.; Djuric, S. W. Topography-biased compound library design: the shape of things to come? *Drug Discovery Today* **2007**, *12*, 948–952.

(43) Among other descriptors, Hajduk et al. found the pocket compactness to be important. However, they defined compactness as the ratio between the volume and the molecular surface area of the binding site. In this equation, with increasing size of the binding site, the numerator will alway increase faster than the denominater. Accordingly, this descriptors rather captures the size of the pocket than its shape.

(44) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(45) Perot, 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.

(46) Nencka, R.; Votruba, I.; Hrebabecky, H.; Jansa, P.; Tloust'ova, E.; Horska, K.; Masojidkova, M.; Holy, A. Discovery of 5-substituted-6-chlorouracils as efficient inhibitors of human thymidine phosphorylase. *J. Med. Chem.* **2007**, *50*, 6016–6023.

(47) Cole, C.; Reigan, P.; Gbaj, A.; Edwards, P. N.; Douglas, K. T.; Stratford, I. J.; Freeman, S.; Jaffar, M. Potential tumor-selective nitroimidazolylmethyluracil prodrug derivatives: inhibitors of the angiogenic enzyme thymidine phosphorylase. *J. Med. Chem.* **2003**, *46*, 207–209.

(48) Neumiller, J. J.; Wood, L.; Campbell, R. K. Dipeptidyl peptidase-4 inhibitors for the treatment of type 2 diabetes mellitus. *Pharmacotherapy* **2010**, *30*, 463–484.

(49) De Clercq, E. Antivirals for the treatment of herpesvirus infections. *J. Antimicrob. Chemother.* **1993**, *32* (Suppl A), 121–132.

(50) Aguero, F.; Al-Lazikani, B.; Aslett, M.; Berriman, M.; Buckner, F. S.; Campbell, R. K.; Carmona, S.; Carruthers, I. M.; Chan, A. W.; Chen, F.; Crowther, G. J.; Doyle, M. A.; Hertz-Fowler, C.; Hopkins, A. L.; McAllister, G.; Nwaka, S.; Overington, J. P.; Pain, A.; Paolini, G. V.; Pieper, U.; Ralph, S. A.; Riechers, A.; Roos, D. S.; Sali, A.; Shanmugam, D.; Suzuki, T.; Van Voorhis, W. C.; Verlinde, C. L. Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat. Rev. Drug Discovery* **2008**, *7*, 900–907.

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

(52) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.* **1998**, *7*, 938–950.

(53) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.