5858 J. Med. Chem. 2010, 53, 5858–5867
DOI: 10.1021/jm100574m

Journal of Medicinal Chemistry Article

Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites

Peter Schmidtke*,‡ and Xavier Barril*,†,‡

†Institució Catalana de Recerca i Estudis Avançats (ICREA), Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain, and ‡Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain

Received May 11, 2010

Druggability predictions are important to avoid intractable targets and to focus drug discovery efforts on sites offering better prospects. However, few druggability prediction tools have been released and none has been extensively tested. Here, a set of druggable and nondruggable cavities has been compiled in a collaborative platform (http://fpocket.sourceforge.net/dcd) that can be used, contributed, and curated by the community. Druggable binding sites are often oversimplified as closed, hydrophobic cavities, but data set analysis reveals that polar groups in druggable binding sites have properties that enable them to play a decisive role in ligand recognition. Finally, the data set has been used in conjunction with the open source fpocket suite to train and validate a logistic model. State of the art performance was achieved for predicting druggability on known binding sites and on virtual screening experiments where druggable pockets are retrieved from a pool of decoys. The algorithm is free, extremely fast, and can effectively be used to automatically sieve through massive collections of structures (http://fpocket.sourceforge.net).

Introduction

Despite advances in both experimental and computational fields, it is estimated that around 60% of drug discovery projects fail because the target is found to be not "druggable".[1] Drug discovery project failures are very expensive, and understanding the difficulties associated with a prospective target is essential to balance investment risks. Since the publication of "the druggable genome"[2] and its estimation of the number of therapeutically useful proteins in the human genome, druggability has gradually become part of the target validation process. Traditional target validation tries to assess whether or not alteration of the normal activity of a potential target can have some significant therapeutic effect. The druggability concept adds a structural dimension and evaluates the likelihood that small drug-like molecules can bind a given target with sufficient potency to alter its activity. Several structure-based druggability prediction methods have been published (reviewed in ref 3). Notable contributions in this domain were first done by the group of Hajduk et al.,[4,5] who used NMR-based fragment screening hit rates as a measure of druggability. The model, based on a simple regression analysis, used descriptors like the surface area, the polar/apolar contact area, the roughness, and the number of charged residues in the binding pocket. In 2007, Cheng et al.[6] published a very simple model to estimate the maximum affinity that an ideal drug-like molecule could have for a given binding site. This model was remarkable because correct predictions were obtained on the assumption that binding affinity of drug-like molecules may derive exclusively from the hydrophobic effect. Drug-target molecular recognition is, nevertheless, a more complex phenomenon,[7] and as our understanding of druggability gradually improves, it will influence our view of the drug binding event, just like pharmacokinetics and drug-likeness have influenced each other.[8,9]

With few exceptions,[10] druggability predictions are based on empirical structure−activity relationships, which require a substantial data set on which to train and validate the model. In that regard, the pioneering studies carried out at Abbott and Pfizer are particularly important because they provided an initial pool of test cases which has facilitated subsequent developments.[11] However, these are still limited and, because the druggabilty concept allows for different interpretations, the classification given to some of the targets is debatable. In this study we unify the previous sets and extend them with further examples, adhering to Cheng's et al. definition of the term "druggable", i.e., capable of binding oral drugs. Undoubtedly, the enormous range of binding affinities and bioavailability rates exhibited by this drug class, as well as the fact that some of these molecules are in fact pro-drugs, introduces a fair amount of ambiguity to the definition. Nevertheless, druggability scores will be extremely useful even if they can only provide a qualitative classification between "druggable", "borderline", and "non-druggable". To facilitate further studies, to promote community involvement in the generation of a larger data set, and to reach a wider consensus on the druggability classification, we have made our test set publicly available and editable (http://fpocket.sourceforge.net/dcd).

Initial studies set the path for druggability predictions, but the resulting algorithms were not made available.[4,6] More recently, Schrödinger have used Cheng's data set to endow their SiteMap cavity detection program with a druggability

*To whom correspondence should be addressed. Phone: +34-934031304 (P.S. or X.B.). Fax: +34-934035987 (P.S. or X.B.). E-mail: pschmidtke@ub.edu (P.S.) or xbarril@ub.edu (X.B.).

 Published on Web 07/12/2010

score (Dscore),[11] but it was trained and presented for usage in specific target validation. Thus, the user has to select a very precise zone for performing the druggability prediction. Because the method was trained to estimate druggability of well-defined cavities, it might not be suitable to assess the multitude of cavities that occur in protein structures, many of which have no reported functional role.[12] Automatic predictions on large structural databases would offer the opportunity to identify druggable cavities on sites or targets that might not be considered a priori, such as allosteric sites or proteins for which a therapeutic rationale has not been fully developed. With this goal in mind, the work presented in this article describes a new structure-based target druggability prediction score coupled to the open source cavity prediction algorithm fpocket.[13] We demonstrate the ability of the method to accurately evaluate the druggability of automatically detected cavities, which allows us to correctly rank binding sites both across and within protein structures. At 2 to 4 s per averaged-sized structure, the method is at least 1 order of magnitude more efficient than SiteMap, and the whole of the Protein Data Bank (PDB[a]) can be processed in a few days on a normal computer. fpocket, including the herein presented druggability score, is freely available for download at http://www.sourceforge.net/projects/fpocket.

Last, we discuss which cavity descriptors are more useful to detect druggable cavities and what they tell us about molecular recognition between proteins and drugs. As shown in previous studies, hydrophobicity correlates particularly well with drug binding sites but the implication of a hydrophobicity-explains-all model such as the MAP_{POD}[6] is that the ideal binding site is a closed and "greasy" cavity. Even in the cavities that more closely resemble this description (e.g., the hormone binding site of nuclear receptors), polar interactions play a fundamental role in binding, selectivity, or mediating the biological response.[14] New evidence is presented here for the special characteristics of the polar groups present in binding sites, which provides a more complex and realistic picture of drug−protein molecular association.

Results and Discussion

Compilation of an Open Data Set. A set of protein−oral drug complexes was obtained crossing the list of marketed oral drugs provided by Vieth et al.[15] with the PDB.[16] The DrugBank[17,18] target information was then used to ensure that the complex corresponds to the actual drug-target pair. Visual inspection ensued to classify the complex as druggable, difficult (e.g., in the case of prodrugs), or undruggable (e.g., in the case of nondruglike ligands). Cheng's data set[6] and the public part of Hajduk's data set[4] were added to obtain the druggability data set (DD) used in this study. As most targets are represented by several structures, and to avoid too much bias toward certain protein families, a nonredundant druggability data set (NRDD) was established using a 70% identity cutoff. The composition of the data sets is summarized in Tables 1 and 2.

The notion of druggability is often riddled with uncertainty and classification can be difficult and may evolve over time. A good example of that is the different classification given by Cheng et al. to the related serine proteinases thrombin and factor Xa as "difficult" and "druggable", respectively.[6] As discussed by Halgren, they share common characteristics and it is likely that, eventually, thrombin will become "druggable" although it is an objectively difficult target.[11] For this reason, we deemed necessary to leave the classification of targets open for discussion. Furthermore, we make the data set public in an attempt to instigate participation of scientists from the field into the creation and design of a unified data set. As demonstrated in the docking arena, the establishment of general benchmarks is important to ensure a fair evaluation of prediction performance.[19,20] In the case of druggability predictions, some targets are more easily predicted than others and, even for the same protein, classification may depend on the particular conformation adopted by the receptor. The use of a common data set will facilitate further developments in the field and avoid biases when comparing methodologies. The Druggable Cavity

[a] Abbreviations: ASA, accessible surface area; DD, druggability data set; DCD, druggable cavity directory; MOc, mutual overlap criterion; NRDD, nonredundant druggability data set; PDB, Protein Data Bank; ROC, receiver operator characteristics.

Table 1. Composition of the DD Data Set; For Comparison, The Composition of the Cheng and Hajduk Data Sets Are Also Shown

	no. of protein structures					
	NRDD[a]	DD[b]			Cheng[c]	Hajduk[d]
druggability classification	holo	holo	apo	total	holo	holo
druggable	45	773	146	919	17 (43)	35
nondruggable	20	75	9	84	4 (10)	37
prodrug	5	60	7	67	6 (10)	
total	70	908	162	1070	27 (63)	72

[a] Nonredundant data set, one structure per protein. [b] Total number of structures in the druggability data set. [c] Reference 6 number of proteins and, in brackets, number of structures. [d] Reference 4 it contains only one structure per protein.

Table 2. Distribution of Structures in DD per Function of the Protein, As Annotated in Uniprot

druggability	no. of structures	function	average resolution
druggable	170	lyase	1.76
	150	hydrolase	1.97
	149	kinase	2.02
	138	nuclear hormone receptor	2.06
	123	oxidoreductase	1.76
	55	Hy[a], Pr[b]	1.98
	42	Hy[a], Pr[b], Tr[c]	2.49
	41	Tr[c], Ox[f]	2.08
	6	structural protein	2.85
	6	transport protein	1.84
	4	ligase	2.17
	4	transferase	2.30
	3	isomerase	1.78
	3	GPCR	3.20
nondruggable	42	Hy[a], Pr[b]	2.12
	21	hydrolase	1.95
	11	transferase	2.32
	4	Lyase	2.18
	3	Tr[c], K_i[d]	2.03
	3	oxidoreductase	1.98
	1	Hy[a], Pr[b], Tr[c], Po[e]	2.10
prodrug binding	39	Hy[a], Pr[b]	2.05
	22	hydrolase	1.85
	3	oxidoreductase	2.82
	3	penicillin binding TMP	2.73

[a] Hydrolase. [b] Protease. [c] Transferase. [d] Kinase. [e] Polymerase. [f] Oxidoreductase.

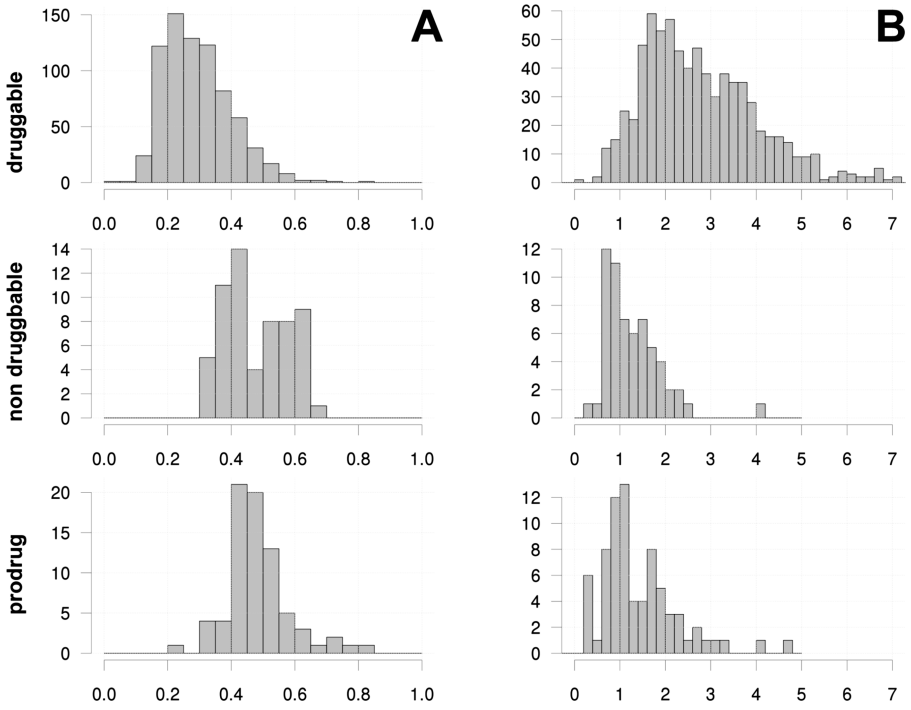

Figure 1. Distribution of (A) the fraction of polar ASA of the pocket and (B) the ASA profile slope ratio on the drugability data set (DD).

Directory (DCD) is a web-based platform that allows multiple users to upload known cavities (identification by PDB code and ligand Hetero Atom Identifier) and assign a value to their druggability in an arbitrary scale from 1 (not druggable) to 10 (druggable). These uploads are then validated by a set of experts (validators) in the field in order to be part of the final data set. Only registered users can upload/validate data. However, anonymous users can access and download the validated data set. The PDB structural database is growing steadily, thus we highly encourage all contributors of the field (medicinal chemists, structural biologists, molecular modellers, etc.) to register, participate, and use this platform, allowing more robust future training/validation cycles for upcoming methods or for retraining of existing methods and scores. The project is available at http://fpocket.sourceforge.net/dcd.

On the Role of Polar Atoms in Druggable Binding Sites. Models with predictive capacity may inform about the physicochemical basis of the underlying process. A good example at hand is the famous rule of five, which predicts drug-likeness based on descriptors related to pharmacokinetics (and very particularly to passive membrane permeation).[21] Similarly, it would be desirable that construction of druggability models could help identify the basis of molecular recognition of drugs by their targets. The Cheng model tells us that the essential feature of a drug binding site is that it should be closed and lipophilic. This was justified on the basis that electrostatic interaction and desolvation energies act in opposition, and the combination of the two is expected to make an insubstantial contribution in the case of charged or polar groups.[6] Nevertheless, the contribution of polar interactions is context dependent, and a single hydrogen bond can contribute as much as 1.8 kcal/mol,[22] comparable to the hydrophobic gain provided by the side-chain of a Val residue.[23] The same applies to ionic interactions.[24] The fact that polar groups play a fundamental role in binding affinity is also supported by the observation that they often constitute anchoring points, featuring predominantly in pharmacophoric models of binding sites.[25] Furthermore, potency is just one of the many factors required for a drug-target complex to result in biological activity. Drugs also have to recognize their target with certain specificity and maintain a stable and specific 3D arrangement within the binding site to be effective. These properties are typically associated with polar interactions, and it seems reasonable to expect that polar groups in binding sites should have some differential properties with regard to the rest of the protein surface. The fact that polar groups are considered irrelevant in Cheng's model or have a negative contribution to druggability in the case of Hajduk et al.[4,6] may wrongly lead to the notion that the ideal druggable site is a completely hydrophobic cavity. We have searched for descriptors with predictive capacity, with the aim of getting an insight on the underlying principles of drug recognition by their targets.

In agreement with previous reports, a predominantly lipophilic composition of druggable binding sites is confirmed in this bigger data set. They typically contain only 20−40% of polar surface versus 40−60% for nondruggable cavities (Figure 1A). Focusing on polar atoms, we find that, on average, 70% of them have very small solvent exposed areas (< 10 Å^2), whereas in nondruggable cavities the proportion decreases to 50%. Considering together the small solvent exposed area of polar atoms and the preponderance of nonpolar atoms, it becomes evident that, in druggable binding sites, protein−ligand hydrogen bonds are surrounded by a hydrophobic environment. In such low dielectric medium, electrostatic interactions become stronger. Very recently, this effect has been quantified in proteins, demonstrating that hydrogen bonds can be up to 1.2 kcal/mol stronger in hydrophobic environments.[26] This clearly indicates that, beyond the obvious gain in hydrophobic potential, a decrease in the polar surface ratio can have the paradoxical effect of increasing the hydrogen bonding potential of the binding site. Without diminishing the importance of hydrophobic interactions, this

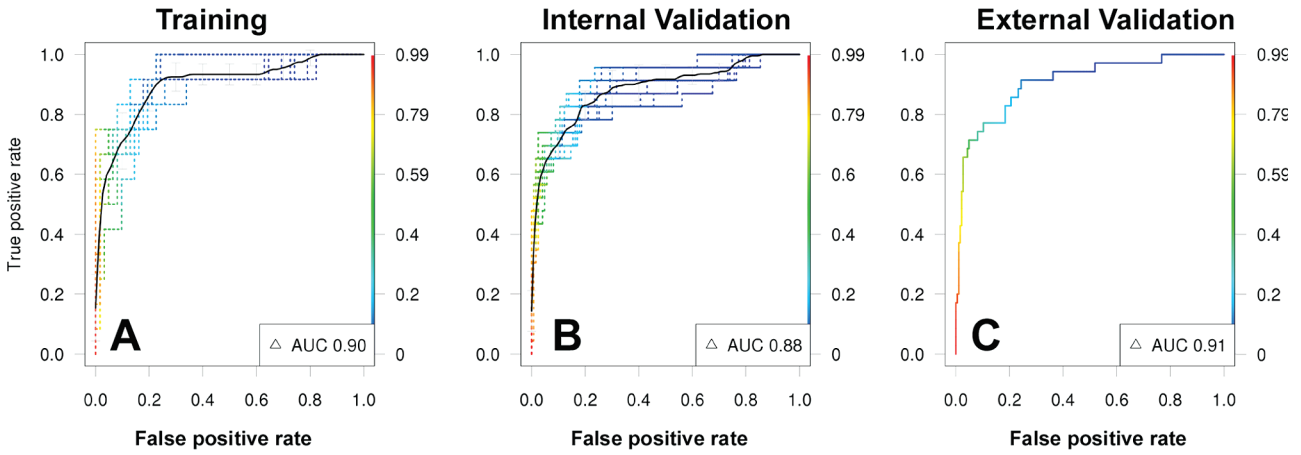

Figure 2. ROC curves for the model building. (A) Training on a total of 74 cavities, (B) internal validation on a total of 146 cavities, (C) external validation on a total of 220 cavities. AUC: area under the curve.

view marries better with the common perception of hydrogen bonds as key elements in drug−protein binding.

Among other descriptors, we investigated the change in accessible surface area (ASA) as a function of the radii used to represent the atoms (see Materials and Methods). Being located in concave regions (cavities), the surface area of a binding site decreases as longer atomic radii are used. Figure 1B shows the ratio of ASA change between polar and nonpolar areas. The different behavior of druggable and nondruggable cavities is highly significant, suggesting that a fundamental aspect of molecular recognition must be associated with this observation. Intriguingly, in nondruggable cavities, the decrease is similar for polar and nonpolar surface areas (average ratio is 1), whereas in druggable cavities, the polar surface area decreases at a much slower rate (average ratio is 3). This means that, in druggable cavities, polar atoms tend to protrude from the cavity surface, making themselves available for interactions (see Figure 6 for a graphical representation). We postulate that the increased protrusion of polar atoms in druggable cavities is fundamental to increase their visibility, rendering them available for interactions and extending the range on which they can exert their selective action. These results also prompted us to investigate the effect of the local environment on the energetics of association of polar groups. In a separate paper (Schmidtke et al., in preparation), we show that the type of local environment found in druggable cavities protects the hydrogen bonds formed between the ligand and the receptor, effectively locking the ligand and permitting longer residence times. Kinetic stability (both of the complex and the binding mode) is another fundamental property of protein−drug complexes that cannot be explained simply on the basis of shape and lipophilic interactions.

Druggability Score. A druggability score was trained and validated on the NRDS, following the protocol described in Materials and Methods. The result of a 10-fold bootstrap run on the learning and internal validation sets is depicted in parts A and B of Figure 2, respectively. The mean prediction result is represented as solid black line, which shows good and stable enrichment. Each cross-validation result is represented as score-colored dashed line. Half of the NDRS was set aside as external validation set, on which the performance of the average model resulting from the learning process was tested (Figure 2C). The resulting scoring function is a two-step logistic function represented by eqs 1−3 and parameters in Table 3 (Materials and Methods). It should be noted that the final formula reflects the need to provide robust predictions in spite of the variability introduced by the automated pocket detection algorithm. Although individually informative, the descriptors in Figure 1 did not yield very robust models during the 10-fold bootstrap learning and validation procedure. These ASA-based descriptors gather information on atomic detail, but as the fpocket cavity prediction can be rather variable from one structure to another or from one conformer to another, their corresponding values are not sufficiently consistent. Instead, the most important descriptor in terms of predictive performance is the mean local hydrophobic density of the binding site. This descriptor combines size and spatial distribution of hydrophobic subpockets into a single number. The two other descriptors used to predict druggability are the hydrophobicity and normalized polarity scores, both of which refer to the physicochemical character of the amino acids lining the pocket. It is noteworthy that both contribute favorably to the score, indicating that both hydrophobic and polar residues can make the binding site more druggable. Again, the residue-based character of theses descriptors makes them more granular, but they also show far less variability than atom-based descriptors during the construction of the final scoring function. As one of the main aims of this scoring function is its high-throughput applicability, it must deliver robust predictions in spite of the varying definition of the binding site provided by the automatic pocket detection algorithm. Thus we would like to emphasize that the scoring function reflects the application for which it has been designed. Descriptors such as those described in Figure 1 are more interpretable and provide a finer level of detail but may be better suited for supervised applications such as predicting the druggability of known binding sites.

As the score is indented for high-throughput and fully automatic predictions, it is necessary to assess its robustness across different crystal structures of the same target. With that purpose, the scoring function established with the NRDS was applied to the whole of the DD. This is a major difference with previous methods. Hajduk et al.[4] used a single structure per protein, whereas Cheng et al.[6] used a variable number of structures, but with no apparent logic. For instance, p38 MAP kinase, for which multitude of structures are available in the PDB, is represented by a single structure (1KV1; DFG-out conformation). Figure 3 illustrates the average score and the standard deviation for each target in the data set. Satisfactorily,

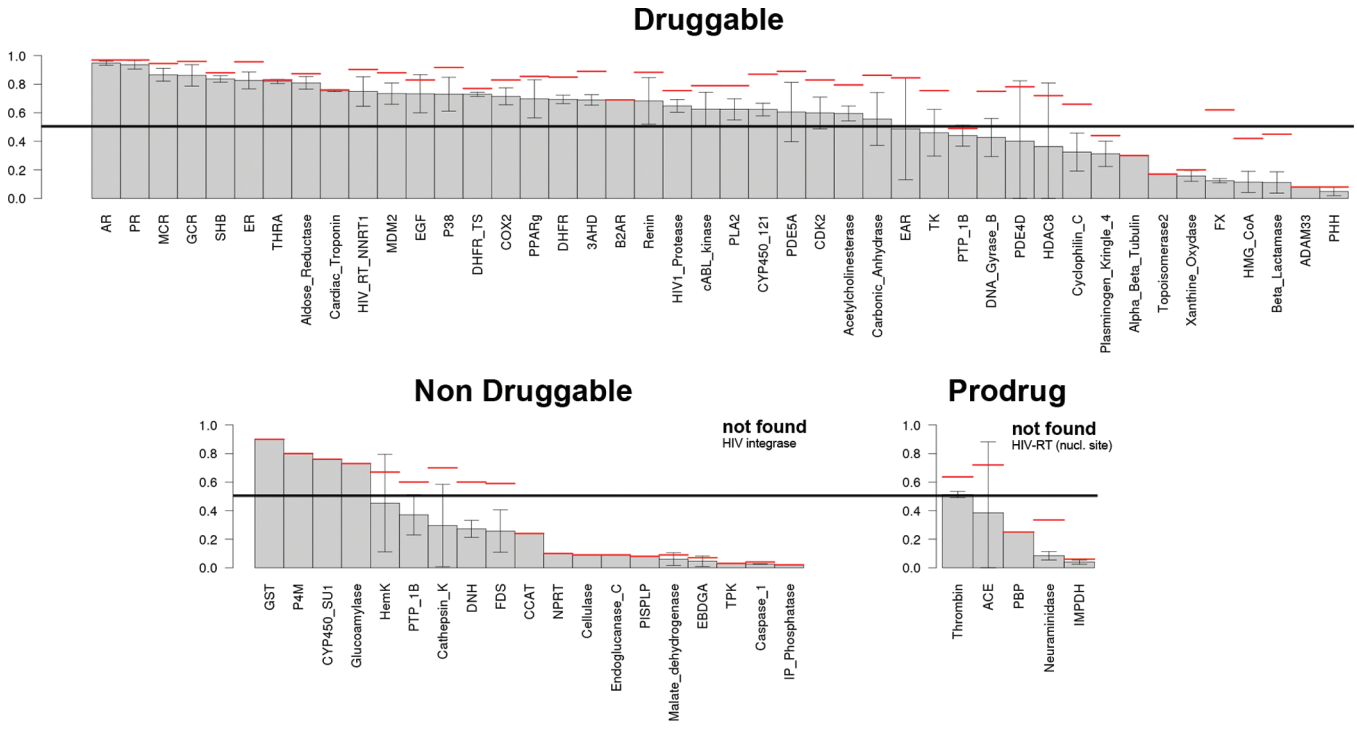

Figure 3. Prediction of druggability on all structures of the DD. Error bars correspond to mean prediction ± standard deviation. For details and full list of protein name abbreviations, refer to Supporting Information.

the standard deviation is low at both ends of the distribution. Variability is larger for scores approximating 0.5, but this is a natural consequence of using a logistic model. Another cause for variability is the pocket prediction itself. In the case of buried and fairly rigid cavities such as the nuclear hormone receptors, the pocket detection algorithm produces consistent results and the druggability score is very stable. In solvent exposed or flexible cavities, the automated pocket detection may yield significantly different binding site predictions and the druggability score may diverge. HDAC8 or CDK2 exemplify this situation, as manifested by very large standard deviations and individual predictions ranging from nondruggable to very druggable. For instance, the average drug score for HDAC8 is 0.36, but the best scoring cavity gets a value of 0.72, comfortably within the druggable range. On rare occasions, the pocket detection algorithm may completely fail to identify the drug binding site. On ACE, fpocket detects the whole internal channel system as one single and continuous cavity, thus it is too large (up to 7700A^3) to consider it as a proper definition of the binding site. For very shallow binding sites, such as the HIV integrase, a cavity may not be detected at all. Arguably, this is not a bad result but a mere reflection of the nondruggable character of the site. In summary, variability in the cavity definition step is an intrinsic limitation of the method presented here, but constraining the druggability prediction to a very concise zone around the experimentally known binding pocket would forbid large scale applicability of the method. Notwithstanding this limitation, the results in Figure 3 and Table S1 (Supporting Information) demonstrate the predictive performance of druggability measurements, indicating that fpocket most often produces consistent results and that the drug score formula manages well pocket variability.

The fact that information in the PDB is very often redundant is an advantage for the method, as multiple predictions can be obtained for a given site. Druggability predictions can then be based on the average score (with a cutoff of 0.5), but if the associated standard deviation is large, it may be advisible to take the value for the top scorers instead. Inspection of Figure 3 (red lines) indicates that, in this case, a cutoff value of 0.7 provides better discriminating power between the druggable and nondruggable sets.

Protein flexibility may also be a source of variability. For instance, the PDE4D binding site can be exposed to solvent or it may be closed due to interactions between the UCR2 domain and the catalytic domain. Interestingly, this conformational change may influence druggability, as a recent paper[27] shows that allosteric modulators binding to the former state have improved side effects compared to known PDE4D inhibitors. All binding sites containing the allosteric modulator were scored (six pockets originally not included in the test set), yielding a higher and less variable drug score than the open conformations (0.55 ± 0.12 and 0.4 ± 0.28, respectively). Another system, not originally included in the druggability data set, for which conformational changes have been associated to different degrees of druggability is renin.[28] We have therefore calculated the druggability on structure 2BKS, which contains two monomers. Chain A has no inhibitor and adopts the conformation that, as stated by Davis et al., is less suitable for drug discovery.[28] Nevertheless, it should be noted that aliskiren is an approved drug targeting this conformation.[29] On chain B, ligand binding induces opening of a hydrophobic subpocket (Figure 4), allegedly making it more druggable. The score obtained was 0.9 and 0.93, respectively. So in this case the structural rearrangements do not alter significantly the druggability prediction, which suggests that the difficulty to develop drugs from the closed conformation may be related to the chemical scaffold of the inhibitors (peptidomimetics) rather than to the physical−chemical properties of the binding site.

Another protein yielding variable predictions is the druggable carbonic anhydrase. However, here the prediction error is due to the fact that many structures contain twice the same ligand, one in the actual binding site and a second

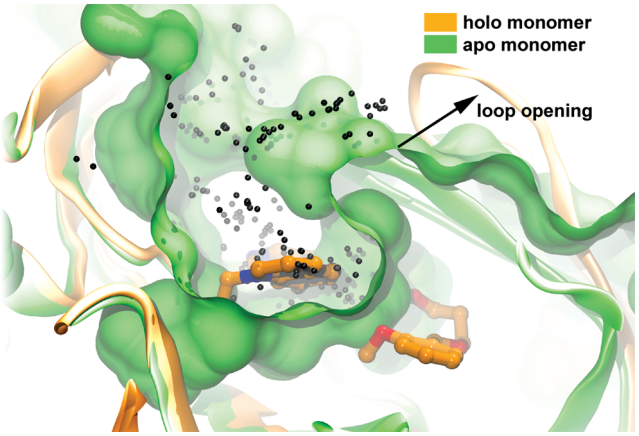

Figure 4. Renin druggable binding site yielding similar druggability scores for holo and apo monomers.

one in a more superficial cavity (e.g., PDB codes 2QOA, 2NNS). As the ligand is used to identify the binding pocket, two completely different binding sites are thus scored. Satisfactorily, they yield opposing drug scores. Accordingly, the variability seen on carbonic anhydrase is not a limitation of the method but of the data set used to evaluate the method.

Comparison with MAP_{POD}. Analyzing the results obtained for the structures in the Cheng data set, few noticeable differences can be observed with the MAP_{POD} score.[6] Proteins like HIV RT (NNRTI site), COX2, CDK2, MDM2, CYP450_121, EGF, PDE5A, acetylcholinesterase, p38, and others are clearly classified as druggable by both methods. Nevertheless, the scores of one protein relative to another do not correlate between models, reflecting the fact that our score is a binary classifier whereas the MAP_{POD} value aims at predicting the maximal binding affinity for the binding site. Agreement is also obtained for the nondruggable proteins cathepsin K, caspase 1 (ICE-1) and PTP-1B, which yield rather low druggability scores. As mentioned above, HIV Integrase has a very shallow binding site that is not even identified by fpocket, indicating that the binding site is not buried enough to wrap a small molecule.

Regarding proteins binding prodrugs, results show a clear separation between druggable binding sites and such "difficult" to target binding sites, again in consonance with the MAP_{POD} score. Considering that druggability score was trained as a bimodal predictive model and none of the prodrug binding sites was used during training or validation, classification of proteins in this category in the nondruggable class is a desirable behavior of the model. Particularly encouraging is the case of thrombin, which receives a score of 0.5. As discussed by Halgren,[11] dabigatran etexilate, a prodrug targeting thrombin, is now marketed in Europe and Canada. Approval by the U.S. Food and Drug Administration is pending for this year.

There is an apparent discrepancy between druggability score and MAP_{POD} on proteins Factor Xa and HMG-CoA reductase. But, in fact, both proteins also yield rather low MAP_{POD} scores, 100-fold lower than the following protein in the druggable data set (DNA gyrase B). The low scores obtained by both methods on HMG-CoA reductase can be explained by the very polar nature of the binding site, which forms ionic interactions with the drugs (e.g., rosuvastatin). Factor Xa, on the other hand, has a heterogeneous and partly shallow cavity, which could influence the low MAP_{POD} score. In the present study, the low score results from the binding site identification protocol, which yields two distinct pockets instead of one.

In conclusion, the method presented here is able to reproduce results obtained by previous publications while not focusing on the binding site of interest, which is a crucial requirement for automated high-throughput druggability predictions.

Wrong Predictions or Wrong Druggability Status? Sometimes, target misclassification can be directly attributed to the ambiguous nature of the "druggability" concept. In the case of druggable cavities with very low score, the pockets are usually very small or host ionic interaction patterns (P-hydroxybenzoate hydrolase, ADAM33). In other cases, the protein forms covalent bonds with the ligand (β lactamase and xanthine oxidase). In those cases, druggability is largely the result of a specific chemical feature rather than a global property of the cavity. Correct assignment of these cavities may therefore require a completely different approach to the one used here. For large-scale predictions, these failures should not be significant, as the sites correspond to enzymatic catalytic centers whose relevance is already evident and can be detected by other means.[30] A borderline druggability score for the allosteric binding site of PTP-1B can hardly be considered a failure given that these inhibitors are weak[31] and the binding site offers limited opportunities for tight binding.[10]

Good druggability scores for nondruggable targets GST, P4M, cytochrome P450 105A1, and glucoamylase, can be traced to the fact that all of these proteins do have well-defined binding sites. They were initially selected because they are listed in the DrugBank as targets of an approved drug and the drug−target complex is available in the PDB. Upon detailed analysis, it was decided that the targets could not really be considered as druggable and were included in the nondruggable data set with the aim of improving the balance between positive and negative data. Reasons to reverse the original classification included wrong drug-target assignment in the DrugBank (e.g., P4M was listed as a target of levodopa), lack of drug-likeness on the ligand part (e.g., the GST inhibitor ethacrynic acid), or the biological role (e.g., GST and CYP are detoxifying enzymes). The druggability status of these proteins is therefore debatable. Glucoamylase is known to bind the oral drug acarbose with high affinity,[32,33] but because this drug acts in the intestine and is not bioavailable, the original classification as nondruggable by Hajduk et al.[34] seems adequate. In all these cases, SiteMap also predicts the sites as druggable or difficult, never as nondruggable. The DCD provides an adequate environment to reconsider the classification of these and other proteins in the data set.

Druggability Prediction on Apo Structures. Because of induced-fit effects, holo structures may present a different arrangement of the features that determine binding. In the molecular docking field, holo structures have been shown to provide qualitatively better results.[35] As training and druggability predictions have been carried out on holo structures, it was necessary to test the robustness of predictions on apo structures to ensure that unused druggable cavities can be found when screening large structural databases (Figure 4). As shown on Table S1 (Supporting Information), the druggability score observed in holo structures is generally reproduced within the range of confidence in apo structures. The main source of variability appears to be the number of apo

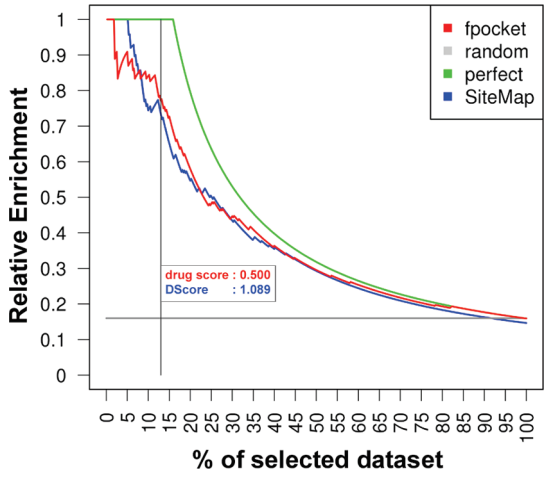

Figure 5. Comparison of druggability prediction performance between fpocket druggability score and SiteMap Dscore.

structures, which is comparatively small. In fact, there are only two proteins with a large number of apo structures: carbonic anhydrase and β lactamase. The latter is predicted as nondruggable, and the average value is identical for the apo and holo sets (0.11). The former is interesting because it is the most populated set (83 holo and 61 apo structures), and the results suggests that there is some information decay as the apo structures get worst average values than holo (0.4 ± 0.3 and 0.5 ± 0.3 respectively). Nevertheless, the best apo cavities get druggability scores as high as the best holo structures (close to 0.9). In the rest of the systems, the number of apo structures is insufficient to get statistics, but the individual values fall within the range observed with holo structures. β-2-Aadrenergic receptor/T4 lysozyme chimera is an exception to this rule, as the apo structures (PDB codes: 2R4R, 2R4S) have missing residues on the extracellular side of the transmembrane helices, where the catecholamines bind. This demonstrates that, except in the case of grossly different cavity shapes, the predictions are sufficiently robust to detect apo as well as holo binding sites.

Comparison with SiteMap. Finally, the method presented here was compared with Dscore, a recent contribution by Thomas Halgren from Schrödinger,[11] which is, to the best of our knowledge, the only other software available to screen for druggable cavities out of the box. The performance of both scores was assessed on the druggable and nondruggable structures from the NRDD. The druggability score is intimately linked to the procedure used to define the cavity; therefore cavities are not interchangeable between programs and have to be generated independently. SiteMap analysis was run on 63 structures, as the remaining eight gave problems when running SiteMap in an automated way. Fixing these structures would have required manual intervention, but this was not done because the goal was to simulate an automated large-scale screen for druggable cavities. As a result, the number of cavities differs between methods (70 druggable, 440 total for fpocket; 63 druggable, 430 total for SiteMap). The performance metric used to compare the methods must take into account the different composition of the data set. We have used a normalized enrichment factor, defined as the ratio of druggable cavities in a given subset of the library. Figure 5 plots this value versus the amount of selected cavities after ordering them by decreasing druggability score/DScore. In both cases, the enrichment factor is ideal in the beginning and remains very good throughout. Predictive power of the druggability score decreases sharply after 0.5, which is the expected behavior, as this value corresponds to the inflection point in the logistic model. The corresponding Dscore at the same fraction of the library is 1.1, coinciding with the average value for druggable cavities in the Dscore training set.[11]

In terms of performance, both methods are similarly capable of retrieving druggable cavities from structural databases. Nevertheless, fpocket present two important advantages for large scale screening purposes. First, the method is very fast (1−2 s for structures up to 450 residues compared to a few minutes for SiteMap). Second, fpocket is completely automatic and does not need protein preparation or selection of parameters. Additionally, the logistic scoring scheme provides a natural cutoff for acceptable enrichment values.

Conclusions

Considerations about druggability are becoming part of the target selection process (see, for instance, ref 36). If the structure is available, this can be done by visual inspection of the binding site but, in the absence of clear guidelines, the decision may be largely subjective. Compiling a large set of targets with their associated druggability is an efficient way of making sure that previous knowledge is retained, thus contributing to our understanding of the fundamental processes behind druggability. Here we present the largest druggability data set to date, which the community can freely download, edit, or extend. In comparing druggable to nondruggable binding sites we find that, contrary to previous models,[6] hydrophobicity is not the sole determinant of target druggability, as polar groups also play an important role on the recognition of drug-like molecules. The data set can also be used to train computational methods or to assess their performance. Previously, computational methods were trained to predict the druggability of known binding sites. Here we have placed particular emphasis on the ability of the program to automatically detect binding sites and subsequently assess their druggability. The resulting software is state of the art in terms of druggability prediction performance while having the advantage of being free of charge, open source, and computationally very efficient. As the druggability prediction is directly associated to a cavity detection method, screening for druggable cavities in large structural data sets is straightforward. Application of this method to the PDB could, then, provide insights into the druggable targetome already contained in the structural proteome. Putative drug binding sites can then be further analyzed by complementary methods.[3,10,37] We expect that this will be a useful approach to unlock promising yet largely unpursued mechanisms of action such as allosteric modulation,[38] protein−protein inhibitors,[39] pharmacological chaperones,[40] or interfacial inhibitors.[41]

Materials and Methods

Data Set. Currently, one data set for assessment of druggability prediction methods was commonly used. This data set was provided by Cheng et al.[6] and was further used for validation of SiteMap druggability score.[11] For a more robust validation of this method, a bigger and nonredundant data set is provided. The herein presented data set was derived using a study published in 2004 by Vieth et al.[15] In this paper, characteristic physical properties and structural fragments of marketed oral

drugs were derived from an extensive data set. Only orally available marketed drugs were kept from this data set.

Next, PubChem (pubchem.ncbi.nlm.nih.gov) was used to check whether a 3D structure of the drug in whatever protein exists in the PDB.[16] Only drugs having resolved 3D structures were kept. Next, data was crossed with DrugBank[17,18] entries for these drugs. The DrugBank contains entries of the targets corresponding to each drug. The known 3D structures from PubChem Compound, corresponding to the actual target of the drug, were kept for further analysis. Finally, structures were checked by hand to establish whether the drug in the protein could be classified as drug, prodrug, or if the binding site should be considered undruggable due to missing drug likeness of the ligand. Structures from Cheng's data set were also added, and the same set was enhanced by other structures for the same proteins, resulting in the druggability data set (DD). Generally, crystal structures with a resolution lower than 2.5 Å and R_{free} below 0.3 were kept despite some exceptions for underrepresented protein classes.

These steps allowed building up of an extensive data set containing for major parts druggable proteins. However, known negative information is also very important and very difficult to find in this field. To enhance the data set with known negative data, the nondruggable proteins from Cheng's data set and parts of the data set published by Hajduk and co-workers[4] was used. Table S1 in Supporting Information summarizes the contents of the data set as well as the prediction results.

Last, for druggable, prodrug binding and nondruggable proteins, a nonredundant data set (NRDD) was established using the BlastClust clusters from the PDB at a maximum of 70% sequence similarity (ftp://ftp.wwpdb.org/pub/pdb/derived_data/NR/clusters70.txt). For learning and validation, only well-defined known druggable and nondruggable cavities were chosen. This was done using the fpocket mutual overlap criterion (MOc),[13] that allows assessment if a found cavity covers well the actual ligand binding site or not. Thus, known druggable and nondruggable cavities needed a MOc of 1 in order to be considered for learning and validation of a druggability scoring function.

Cavity Detection. For this study, fpocket, a highly scalable and free open source pocket detection software package, was used.[13] Extensive usage was made especially of the dpocket program, allowing easy extraction of pocket descriptors. The default set of dpocket descriptors was extended by polar and apolar pocket surface area (van der Waals surface +1.4 Å and van der Waals surface +2.2 Å). The dpocket derived pocket descriptors were further tested for suitability in the creation of a druggability score using logistic regression.

Finally, the retained pocket descriptors are:

- The normalized mean local hydrophobic density. This descriptor tries to identify if the binding pocket contains local parts that are rather hydrophobic. For each apolar α sphere the number of apolar α sphere neighbors is detected by seeking for overlapping apolar α spheres. The sum of all apolar α sphere neighbors is divided by the total number of apolar α spheres in the pocket. Last, this score is normalized compared to other binding pockets on the same protein.
- The hydrophobicity score. This descriptor is based on a residue based hydrophobicity scale published by Monera et al.[42] For all residues implicated in the binding site, the mean hydrophobicity score is calculated and is used as descriptor for the whole pocket. Each residue is evaluated only once.
- The normalized polarity score. As published on http://www.info.univ-angers.fr/~gh/Idas/proprietes.htm, each residue can be split in two polarity categories (1 and 2) The final polarity score is the mean of all polarity scores of all residues in the binding pocket. Each residue is evaluated only once.

Table 3. Constants of Druggability Score Model

descriptor	coefficient	mean value[a]	standard deviation/mean[b]
intercept	ss_0	−6.238	−0.095
mean local hydrophobic density (normalized)	ss_1	4.592	0.154
hydrophobicity score	ss_2	5.717	0.170
polarity score (normalized)	ss_3	3.985	0.459
intercept	$ss_{1,0}$	−5.141	−0.170
mean local hydrophobic density (normalized)	$ss_{1,1}$	6.579	0.173
intercept	$ss_{2,0}$	−2.669	−0.168
hydrophobicity score	$ss_{2,1}$	0.056	0.216
intercept	$ss_{3,0}$	−2.445	−0.238
polarity score (normalized)	$ss_{3,1}$	2.762	0.330

[a] Mean values refer to mean constants after a 10-fold bootstrap run. [b] The ratio between the standard deviation and the mean value of the constant assess the variability of the constant during the bootstrap.

Druggability Score. As the NRDD contains only 45 druggable and 21 nondruggable proteins, the following rule was considered: all other cavities (not in contact with a ligand) identified on the proteins containing at least one druggable cavity and having a size higher than 60 α spheres (corresponding to reasonably sized cavities) were also considered as nondruggable. This rule allows to introduce decoys into the given set of cavities in the NRDD and thus increase its size substantially.

To train and validate the druggability score, the NRDD, consisting of 70 druggable cavities, 16 nondruggable cavities having a MOc of 1, and 354 decoys was split in two. The first half of the data was used to train the model. Training and internal validation was performed using a 10-fold bootstrap with a one-third/two-third training/validation ratio. First logistic models were derived for each pocket descriptor using the glm function from the R statistical software package,[43] and according to predictive power and stability during the 10-fold bootstrap, these models were further considered.

In the next step, predictions coming from these "one descriptor based" logistic models were associated in one common logistic model where statistically nonsignificant descriptors or unstable models were filtered out. The general model is shown in eqs 1−3 as drugscore, the single descriptor based models are designated by the function $f_x(d_x)$, where d_x is a given descriptor.

$$\text{drugscore}(z) = \frac{e^{-z}}{1+e^{-z}} \tag{1}$$

$$z = \beta_0 + \beta_1 f_1(d_1) + \beta_2 f_2(d_2) + \beta_3 f_3(d_3) \tag{2}$$

$$f_x(\mathrm{d}_x) = \frac{e^{-\beta_{x,0} + \beta_{x,1} d_x}}{1+e^{-\beta_{x,0} + \beta_{x,1} d_x}} \tag{3}$$

The coefficients of the model, as shown in Table 3, were obtained by averaging coefficients derived on each step of the bootstrap.

The second half of NRDD was reserved for external validation (selected by random before the bootstrap run). Receiver operator characteristics (ROC) and derivative figures shown in this paper were produced using the ROCR package.[44]

Comparison with Schrödinger SiteMap. To compare fpocket druggability prediction performance with SiteMap Dscore, the NRDD data set was used as benchmark. SiteMap was systematically launched on all structures after running the Prepwizard protein preparation protocol of Maestro. SiteMap was run to accept a maximum number of 10 binding sites to reproduce prediction results published by Halgren. However, no binding site restrictions were applied. A binding site was successfully recognized if at least 20 site points were less than 1.5 Å away

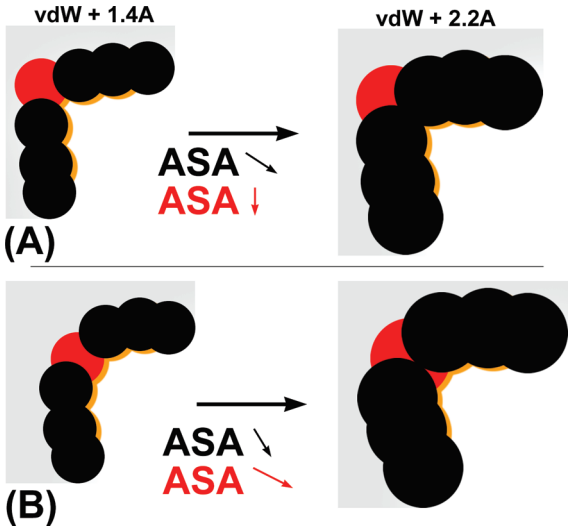

Figure 6. Principles of ASA calculation upon atom fattening. Apolar atoms are colored in black, polar in red, the ASA is shown in orange: (A) a buried polar atom has a low ASA when taking into account the van der Waals radius +1.4 Å. This ASA disappears if one increases the probe size to 2.2 Å. (B) A more exposed polar atom has also a low ASA when taking a van der Waals radius +1.4 Å but has still a contribution to the total ASA of the pocket when taking the van der Waals raidus +2.2 Å.

from any of the atoms of the known ligand. Unlike the training/validation procedure, all other binding sites on druggable and nondruggable proteins were considered as decoys for both algorithms.

For this comparison, fpocket was run on the NRDD and the druggability score calculated for each binding site. A binding site was considered druggable/nondruggable according the rules specified in the previous part of Materials and Methods.

Pocket Surface Calculations. To assess the importance of polar atoms in known drug binding sites (bound structure), surface calculations were performed in the following way. The set of pocket atoms were identified as the atoms within at most 5.5 Å from the nearest ligand atom. For this set of atoms, the portion of van der Waals surface nonoccluded by surrounding atoms was calculated.

To calculate an ASA close to the solvent-accessible surface area, the van der Waals radius of each pocket atom was increased by 1.4 Å for the surface calculation.

Next, atom fattening was performed, increasing the van der Waals radius correction from 1.4 to 2.2 Å in steps of 0.1 Å. For each modified van der Waals radius, the van der Waals surface of the pocket was calculated. By definition, in concave portions of the protein surface, the van der Waals surface decreases upon atom fattening. In the protein binding sites assessed here, these surfaces decreased linearly. Because of this linear behavior in the radius range considered (van der Waals +1.4 Å to van der Waals +2.2 Å), an automatic construction of linear regression based models is possible and reliable. Thus for the polar and apolar ASA profiles, two models with two parameters (slope and intersect) for each can be obtained. The behavior of these parameters was assessed throughout this study, enabling direct access to the concavity of the pocket for polar and apolar atoms (slope) and the ratio of polar versus apolar atoms (ratio of intersects). The principle is shown in Figure 6. Furthermore, the ratio between the apolar and polar slope of the concavity profiles is of importance in this work and will be referenced to concavity profile ratio. The concavity profile calculation was implemented for running on NVIDIA graphics processing units using python, the excellent Biskit structural bioinformatics framework,[45] and PyCUDA.[46]

Acknowledgment. We thank Vincent Le Guilloux for careful proofreading of the manuscript. This work was financed by the Ministerio de Educación y Ciencia (grant SAF2009-08811). P.S. is funded by the Generalitat de Catalunya. The "Servei de Disseny de Fàrmacs" at CESCA is acknowledged for access to commercial software.

Supporting Information Available: Full list of the results obtained for each protein in the data set. List of protein name abbreviations used in Figure 3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067–1077.
(2) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nature Rev. Drug Discovery* **2002**, *1*, 727–730.
(3) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.
(4) 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.
(5) Hajduk, P. J.; Huth, J. R.; Tse, C. Predicting protein druggability. *Drug Discovery Today* **2005**, *10*, 1675–1682.
(6) 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. *Nature Biotechnol.* **2007**, *25*, 71–75.
(7) Böhm, H. J.; Klebe, G. What can we learn from molecular recognition in protein−ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed. Engl.* **2003**, *35*, 2588–2614.
(8) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 470–477.
(9) Vistoli, G.; Pedretti, A.; Testa, B. Assessing drug-likeness—what are we missing? *Drug Discovery Today* **2008**, *13*, 285–294.
(10) Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principle. *J. Med. Chem.* **2009**, *52*, 2363–2371.
(11) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
(12) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892–906.
(13) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.
(14) Sack, J. S.; Kish, K. F.; Wang, C.; Attar, R. M.; Kiefer, S. E.; An, Y.; Wu, G. Y.; Scheffler, J. E.; Salvati, M. E.; Krystek, S. R., Jr; Weinmann, R.; Einspahr, H. M. Crystallographic structures of the ligand-binding domains of the androgen receptor and its T877A mutant complexed with the natural agonist dihydrotestosterone. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4904–4909.
(15) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224–232.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(17) 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–D672.
(18) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–6.
(19) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(20) 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.
(21) 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.* **2001**, *46*, 3–26.
(22) Fersht, A. R. The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* **1987**, *12*, 301–304.

(23) Karplus, P. A. Hydrophobicity regained. *Protein Sci.* **1997**, *6*, 1302–1307.
(24) Barril, X.; Aleman, C.; Orozco, M.; Luque, F. J. Salt bridge interactions: stability of the ionic and neutral complexes in the gas phase, in solution, and in proteins. *Proteins* **1998**, *32*, 67–79.
(25) Langer, T.; Hoffmann, R. D. Pharmacophores and Pharmacophore Searches. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2006; Vol. 32, pp 375.
(26) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature Struct. Mol. Biol.* **2009**, *16*, 684–690.
(27) Burgin, A. B.; Magnusson, O. T.; Singh, J.; Witte, P.; Staker, B. L.; Bjornsson, J. M.; Thorsteinsdottir, M.; Hrafnsdottir, S.; Hagen, T.; Kiselyov, A. S.; Stewart, L. J.; Gurney, M. E. Design of phosphodiesterase 4D (PDE4D) allosteric modulators for enhancing cognition with improved safety. *Nature Biotechnol.* **2010**, *28*, 63–70.
(28) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem., Int. Ed. Engl.* **2003**, *42*, 2718–2736.
(29) Staessen, J. A.; Li, Y.; Richart, T. Oral renin inhibitors. *Lancet* **2006**, *368*, 1449–1456.
(30) Sankararaman, S.; Sha, F.; Kirsch, J. F.; Jordan, M. I.; Sjolander, K. Active site prediction using evolutionary and structural information. *Bioinformatics* **2010**, *26*, 617–624.
(31) Wiesmann, C.; Barr, K. J.; Kung, J.; Zhu, J.; Erlanson, D. A.; Shen, W.; Fahr, B. J.; Zhong, M.; Taylor, L.; Randal, M.; McDowell, R. S.; Hansen, S. K. Allosteric inhibition of protein tyrosine phosphatase 1B. *Nature Struct. Mol. Biol.* **2004**, *11*, 730–737.
(32) Sierks, M. R.; Svensson, B. Functional roles of the invariant aspartic acid 55, tyrosine 306, and aspartic acid 309 in glucoamylase from *Aspergillus awamori* studied by mutagenesis. *Biochemistry* **1993**, *32*, 1113–1117.
(33) James, J. A.; Lee, B. H. Glucoamylases: microbial sources, industrial applications and molecular biology: a review. *J. Food Biochem.* **1997**, *21*, 1–52.
(34) Andrews, J. S.; Weimar, T.; Frandsen, T. P.; Svensson, B.; Pinto, B. M. Novel Disaccharides Containing Sulfur in the Ring and Nitrogen in the Interglycosidic Linkage. Conformation of Methyl 5′-Thio-4-*N*-alpha-Maltoside Bound to Glucoamylase and Its Activity as a Competitive Inhibitor. *J. Am. Chem. Soc.* **1995**, *117*, 10799–10804.
(35) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265–4272.
(36) Campbell, S. J.; Gaulton, A.; Marshall, J.; Bichko, D.; Martin, S.; Brouwer, C.; Harland, L. Visualizing the drug target landscape. *Drug Discovery Today* **2010**, *15*, 3–15.
(37) Panjkovich, A.; Daura, X. Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. *BMC Struct. Biol.* **2010**, *10*, 9.
(38) Lindsley, J. E.; Rutter, J. Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10533–10535.
(39) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein−protein interactions: progressing towards the dream. *Nature Rev. Drug Discovery* **2004**, *3*, 301–317.
(40) Leandro, P.; Gomes, C. M. Protein misfolding in conformational disorders: rescue of folding defects and chemical chaperoning. *Mini Rev. Med. Chem.* **2008**, *8*, 901–911.
(41) Pommier, Y.; Cherfils, J. Interfacial inhibition of macromolecular interactions: nature's paradigm for drug discovery. *Trends Pharmacol. Sci.* **2005**, *26*, 138–145.
(42) Monera, O. D.; Sereda, T. J.; Zhou, N. E.; Kay, C. M.; Hodges, R. S. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. *J. Pept. Sci.* **1995**, *1*, 319–329.
(43) R Development Core Team. *R: a language and environment for statistical computing*, **2007**.
(44) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940–3941.
(45) Grunberg, R.; Nilges, M.; Leckner, J. Biskit—a software platform for structural bioinformatics. *Bioinformatics* **2007**, *23*, 769–770.
(46) Klöckner, A.; Pinto, N.; Lee, Y.; Catanzaro, B.; Ivanov, P.; Fasih, A. PyCUDA: GPU Run-Time Code Generation for High-Performance Computing. *arXiv* **2009**.