Binding Site Detection and Druggability Index from First Principles

Jesus Seco, F. Javier Luque, and Xavier Barril[†],[*]

Institució Catalana de Recerca i Estudis Avançats (ICREA), Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Spain, Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Barcelona, Spain

In drug discovery, it is essential to identify binding sites on protein surfaces that drug-like molecules could exploit to exert a biological effect. Both X-ray crystallography and NMR experiments have demonstrated that organic solvents bind precisely at these locations. We show that this effect is reproduced using molecular dynamics with a binary solvent. Furthermore, analysis of the simulations give direct access to interaction free energies between the protein and small organic molecules, which can be used to detect binding sites and to predict the maximal affinity that a drug-like molecule could attain for them. On a set of pharmacologically relevant proteins, we obtain good predictions for druggable sites as well as for protein−protein and low affinity binding sites. This is the first druggability index not based on surface descriptors and, being independent of a training set, is particularly indicated to study unconventional targets such as protein−protein interactions or allosteric binding sites.

Introduction

The biological components used as targets in drug therapy meet the double condition of being disease-modifying and having an activity that can be modulated upon binding by small molecules. Target validation, a necessary step before initiating a drug discovery program, aims at clarifying the role played by a putative target in the disease,[1] but until recently little effort was made to ensure that the target could be modulated by a drug-like molecule. Because drug candidates have a greater probability of displaying an adequate pharmacokinetics profile if they are small in size and have a balanced mixture of polar and apolar groups[2,3] and they also have to achieve high binding affinity to exert their action, only those targets providing a maximum level of shape and chemical complementarity to drug-like molecules are effectively useful for pharmacological intervention. Though this notion is widely assumed in drug design,[4] methods to quantify target druggability have only recently emerged.[5,6] Similarly to binding site detection methods, druggability predictions have to be able to identify small-molecule binding sites on a protein, but whereas the former just aims at providing a correct ranking of the sites within a protein, the latter should enable comparison between different targets of interest (e.g., components of a pathway) and provide an estimate of the feasibility of developing a drug candidate, which is essential for risk management and adequate project planning.[7]

A large body of literature indicates that excellent small-molecule binding site predictions can be obtained using surface descriptors (see ref 8 and references therein). The druggability methods proposed by Abbot[5] and Pfizer[6] groups build on this ground to provide druggability indices that are mostly based on the curvature and lipophilicity of the protein surface, although they differ in the definition of "druggability". The index proposed by Hajduk and co-workers is based on the hit rates obtained in NMR-based fragment screening.[5] It relies on the assumption that sites binding a higher proportion of fragments

are also more likely to deliver high-affinity, noncovalent drug-like leads. Thus, the druggability measure is, in fact, a prediction of the fragment screening hit rates. In contrast, the strategy adopted by Cheng et al. pursues prediction of the maximal affinity that a drug-like molecule can achieve for a binding site.[6] This is perhaps a more useful criteria in target validation because the identification of many hits does not guarantee that a drug-like lead should ensue.[9]

Regardless of their differences, both methods are extremely fast and need only an atomic resolution structure of the target to generate the predictions However, as any empirical method, their scope is limited by the composition of the training set, which is inevitably biased toward those target classes that benefit from abundant structural and pharmacological information. For instance, protein−protein interactions are a difficult yet promising target class,[10] which is nearly unrepresented in one of the studies,[6] and accurate predictions cannot be presumed for atypical binding sites. Another potential limitation of the methods is that the predictions are based on a static representation of the target and, although they can be used on ensembles of protein structures,[11] some dynamic properties such as the presence of transient cavities or important interstitial water molecules can be missed. Most important of all, statistically derived structure−activity relationships often fail because the descriptors used to make the predictions do not play a causative role[12] or due to "activity cliffs" (i.e., major changes of activity due to minor structural differences), which are difficult to capture using statistical models.[13] In this context, it should be noted that the underlying cause for binding is the presence of hot spots,[14] and the surface descriptors used to make druggability predictions may not always be good predictors for such privileged binding loci. Indeed, hot spots also exist in flatter surfaces and the published methods would be unable to guess the maximal affinity that a drug-like compound could achieve in such an environment.

Here we present a new druggability index that, being based on first principles, bypasses most of the aforementioned limitations and provides a completely independent estimate. Our approach relies on the evidence (provided by NMR and crystallographic experiments) that binding sites on protein surfaces have a tendency to bind small organic solvents.[15−18]

* To whom correspondence should be addressed. Phone: +34-934029002. Fax: +34-934035987. E-mail: xbarril@ub.edu. Address: Departament de Fisicoquímica, Facultat de Farmàcia, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain.
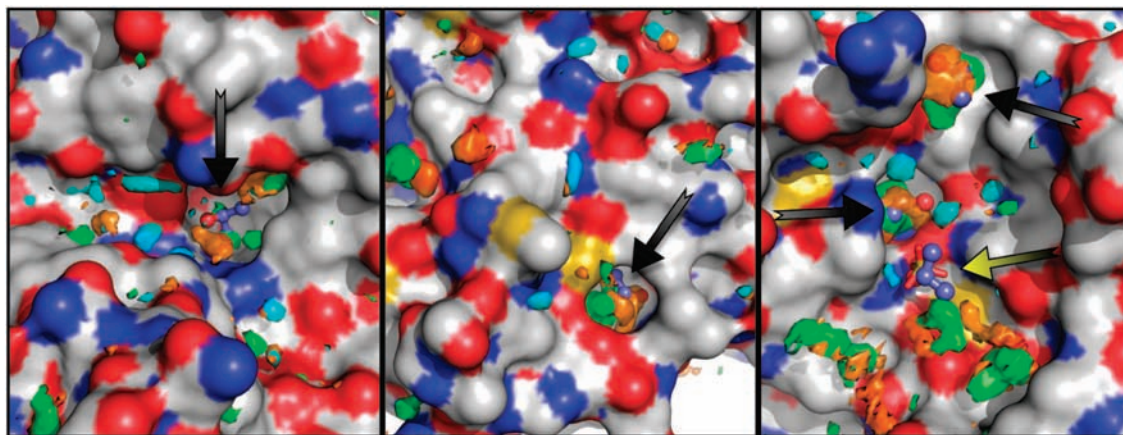† ICREA.

Figure 1. Experimentally determined iPrOH binding sites (arrows) on the surface of thermolysin (left), p53 core domain (center), and elastase (right). These are depicted along high concentration areas for the different solvent types, as extracted from MD simulations. Isosurfaces are color-coded as follows: orange, 16 times the expected density of OH group in iPrOH; green, 16 times the expected density of Me group in iPrOH; cyan, 4 times the expected water density. The yellow arrow signals the oxyanion hole of elastase. Figures 1, 2, 4, and 5 were created with PyMOL v.0.99.[58]

We demonstrate that these distinctive binding features are also reproduced by state-of-the-art computational methods. Furthermore, we show that by analyzing the organic vs water solvation preferences of the protein surface, it becomes possible to detect binding sites and to quantify the maximal binding free energy that a drug-like molecule could achieve for them.

Results

Isopropyl alcohol (iPrOH[a]) is used as a prototypical drug-like molecule as it contains a polar part, featuring both hydrogen-bond donor and acceptor capabilities, and a lipophilic part. To investigate the druggability of a biomacromolecule of known structure, we simply solvate it in silico with an iPrOH/water mixture and perform molecular dynamics (MD) simulations to determine the solvation preferences (see Methods for details). After running the simulations, solvent densities are extracted from the MD trajectories as the count of solvent atoms per volume unit. Given that binding sites, and hot spots in general, have a tendency to become desolvated and interact with small organic molecules,[19] their presence will be revealed by an increase in the local density of iPrOH to a degree proportional to its strength as interaction points. In the following, we first validate the use of MD to sample the solvation preferences of proteins and then we discuss how this approach can be used to detect binding sites and quantify their maximal binding affinity. Finally, we proceed to measure the druggability of a set of pharmacological targets, contrasting the results with experimental data.

Experimental iPrOH Binding Sites. Solvent-mapping studies have been used to locate binding sites on proteins,[20] and they have revealed iPrOH binding sites at the surface of thermolysin,[21] the p53 core domain,[22] and porcine pancreatic elastase.[23] Accordingly, we have used these proteins to test the ability of our strategy to predict the binding site of small molecules. This could, in principle, be achieved with molecular mechanics or docking approaches at a fraction of the computational cost. In fact, multiple copy simultaneous search,[24] computational solvent mapping,[25] and the grand canonical

Monte Carlo method of Locus Pharmaceuticals[26] have been developed for the specific purpose of predicting the binding site of small probe molecules. These methods produce a map of the interaction preferences of binding sites that can be very useful in drug design.[27,28] Nevertheless, they cannot be expected to produce quantitative estimates of binding free energies because: (1) the free energy results from the addition of multiple terms of large magnitude and opposed sign, which typically have large associated errors,[29] and (2) these methods use rather blunt approximations such as treating the protein as a rigid body or ignoring solvation and entropic effects. Using MD simulations with explicit solvent, we do not have to rely on such approximations and, as MD sampling naturally converges to a Boltzmann ensemble, it provides access to free energies without having to calculate each contributing term.

Structures of thermolysin have been obtained at iPrOH concentrations ranging from 2% to 100%, revealing several iPrOH binding sites of different apparent affinities.[21] The highest affinity corresponds to a small pocket on the protein surface, which is already occupied by iPrOH at a concentration of 5%. As shown in Figure 1, left, several areas in this pocket have a population of methyl groups of iPrOH (Me-iPrOH, green isosurfaces) 16-fold larger than the expected value. This also coincides with two other areas similarly populated by the hydroxyl group of iPrOH (O-iPrOH, orange isosurfaces), thus providing an excellent agreement with the experiment. It is also interesting to observe that the rest of the protein surface contains few other sites with high solvent densities, confirming that iPrOH does not bind everywhere on the surface, rather it interacts preferably with particular sites which may be presumed to be binding hot spots. Experimentally, a second binding site appears at 10% iPrOH and other sites are only detectable at concentrations above 60%. The secondary binding site is located in a small cavity in the protein interior with no connection to bulk solvent. Filling this cavity with iPrOH involves, therefore, a partial unfolding and refolding of the protein. As this sort of event happens at timescales much larger than the span of our MD simulation, in this case, the methodology fails to identify the site. Clearly, correct predictions are only possible if solvent molecules can be exchanged at a reasonably fast rate, but we expect that this will be the case for most binding sites on the outer surface of proteins. Of the remaining crystallographic iPrOH molecules, we only observe some density near positions
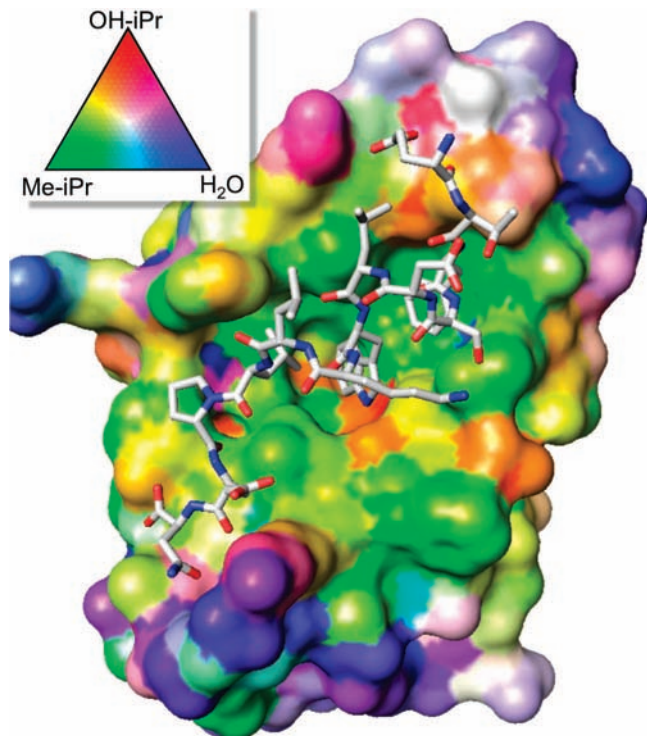
Figure 2. MDM2 surface color-coded according to its solvation preferences. The inset shows the Maxwell triangle, which was used to code interaction preferences: green for Me-iPrOH, red for O-iPrOH, blue for water, and a composite color for mixed preferences. The binding mode of the p53 peptide is shown for illustrative purposes only.

3 and 8 (using the nomenclature used by English et al.), which appear at 80% and 90%, respectively. In both cases, the observed densities are much smaller than for the primary binding site.

In the case of p53 core domain, a soak at 35% iPrOH concentration reveals a single binding site.[22] This is fully reproduced by our calculations, which identify the site as a high density spot both for O-iPrOH and Me-iPrOH (see Figure 1, center). The comparison with experimental results is less straightforward in the case of elastase, as the structure was obtained at a concentration of 80% iPrOH, 4-fold higher than the one used in our simulation.[23] Notwithstanding this difference, two out of the three crystallographic iPrOH molecules are well reproduced by our simulation. The third position (yellow arrow in Figure 1, right) corresponds to the oxyanion hole, which usually binds anions (e.g., sulfate) and is therefore unlikely to bind organic molecules in a high dielectric medium.

Overall, these results indicate that MD with an iPrOH/water mixture can be used to detect the binding sites of iPrOH. These are identifiable as areas where high density of O-iPrOH and Me-iPrOH colocalize. Furthermore, the results also show that this simple technique reveals other patches on the protein surface that are readily dehydrated and prefer to interact with organic molecules. In the next section, we discuss how these putative hot spots can be used to identify binding sites and quantify their druggability.

Density isosurfaces, as plotted in Figure 1, provide a convenient way of visualizing preferential binding sites for polar and hydrophobic groups of organic molecules, as well as for water. Another visually interpretable representation is obtained by projecting the interaction preferences onto the surface of the biomolecule. This is illustrated for protein MDM2 in Figure 2, where the number of contacts made by the protein atoms with each solvent feature (normalized by the expected value) have
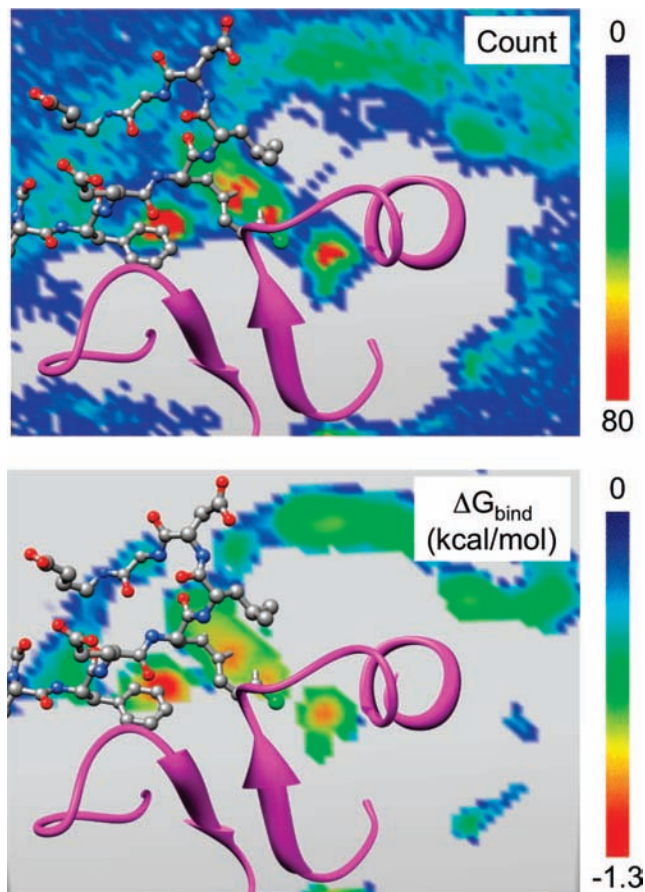

Figure 3. Slice of a grid used to count the number of times that a Me-iPrOH is found into each volume element (top). The same grid converted to free energy of binding using eq 1 (bottom). Image created with Chimera.[59]

been color-coded using a Maxwell triangle. MDM2 binds the N-terminal tail of p53 and is often cited as an example of druggable protein−protein interaction.[30,31] The p53 binding groove is clearly identifiable as a large green area, denoting preference for hydrophobic groups, with an orange−red strip in the center (O-iPrOH binding preference) marking the site where Trp23 of p53 forms a hydrogen bond with the protein backbone. Such graphical representations provide a qualitative indication of the presence of hot spots or binding sites, but a quantitative measure can also be obtained if the solvent follows a Boltzmann distribution.

Detection of Binding Sites and Quantification of Maximal Affinities. After running the simulation, a grid encompassing the whole of the simulation box is generated and the number of times that a solvent feature falls within each grid element is counted. Comparing the observed population (N_i) with the expected value (N_o), the associated free energy can be obtained:

$$\Delta G_i = -k_B T \ln(N_i/N_o) \qquad (1)$$

where k_B is the Boltzmann constant and T the temperature at which the simulation was run.

This is graphically represented in Figure 3, which shows a slice of the grid with the Me-iPrOH count obtained from the MDM2 simulation and its conversion to free energy values. It is worth noting the correspondence between the spots of more favorable interaction free energy and the positions of hydrophobic atoms of a potent MDM2 inhibitor (the cyclic peptide
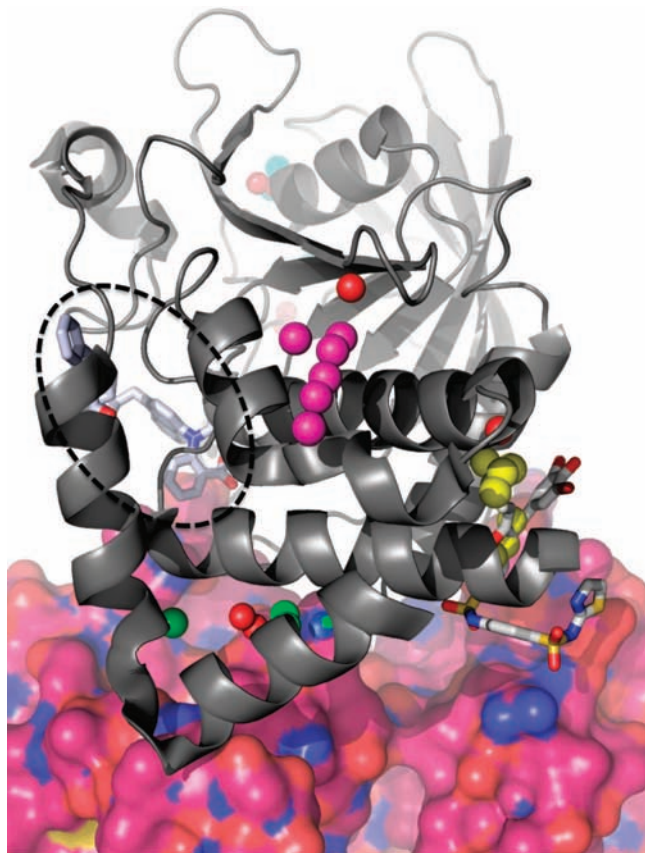
Figure 4. Binding sites detected on the surface of PTP-1B (gray ribbons). The phosphotyrosine binding site, where most ligands bind, has no hot spot for binding (dashed circle). The most potent binding site (green spheres) corresponds to the PTP-1B-IRK interface (IRK shown as surface). The second one (yellow spheres) reveals the binding site of allosteric inhibitors. Two other sites of lower maximal affinity are also shown.

in structure 2AXI[32]). We reasoned that the maximal affinity of a drug-like ligand could be estimated from the free energy grids by applying a few basic rules:

(1) The binding free energy (ΔG_{bind}) of Me-iPrOH and O-iPrOH is transferable to aliphatic and polar neutral features of a drug-like compound, respectively.

(2) An ideal ligand would place atoms of the right type at locations offering optimal ΔG_{bind}, separated between them by distances no shorter than a covalent bond and spread over a volume no larger than a typical drug-like molecule.

(3) Atomic contributions to ΔG_{bind} cannot exceed the -1.5 kcal/mol threshold empirically determined by Kuntz & Kollmann.[33]

(4) Atoms in a ligand play a dual role as part of the chemical framework and as interaction points. As a result, most atoms contribute less to binding than they could.[34] We postulate that the same effect can be reproduced by considering the ligands as a combination of binding features, which make optimal interactions with the receptor's hot spots, and connecting parts, which make negligible contributions to ΔG_{bind}.

We have implemented a protocol (see Methods) that, based on those principles, clusters together several interaction points to form a molecular-shaped volume of maximal binding free energy.

The systems used in the previous section offer a first test set for our method. For p53 core domain, the predicted maximal affinity of a drug-like molecule binding the iPrOH binding site
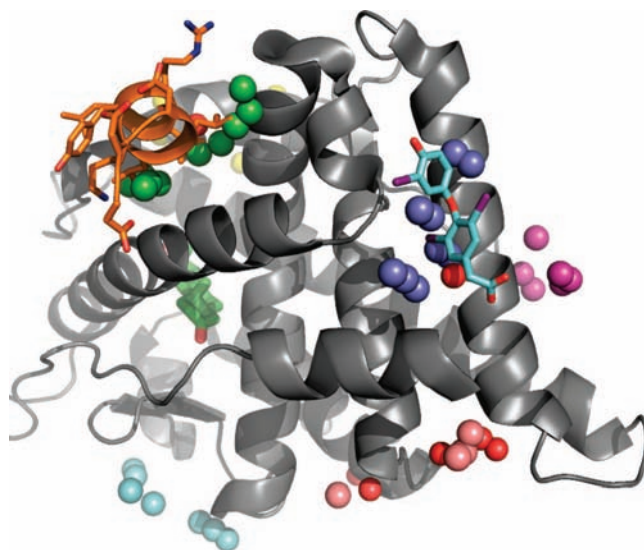

Figure 5. Structure of the androgen receptor with a peptide bound in the AF2 site (shown in orange) and a small organic molecule bound to the BF3 site. These are identified as binding sites with maximal affinities of 1 and 80 nM (green and purple spheres, respectively). Three other binding sites with predicted binding affinities in the range $0.1-10$ μM are also shown.

(and surrounding areas) is only -5.8 kcal/mol, corresponding to a dissociation constant (K_d) of 60 μM, hence it cannot be considered druggable. It should be noted that no functional role has been attributed to this site. The iPrOH binding site of thermolysin, on the contrary, coincides with the catalytic site, which is recognized by several inhibitors. In this case, the maximal affinity prediction is -8.9 kcal/mol ($K_d = 310$nM). Thermolysin inhibitors can be much more potent than that, even reaching sub-nM values, but it should be noted that they all contain an anionic moiety (e.g., carboxylic acid, phosphonate) that forms a complex with the catalytic Zn^{2+}.[35] Considering that metal chelation is very exothermic and that iPrOH does not form contacts with the ion, the prediction seems a reasonable estimate for a molecule containing only lipophilic and polar neutral features. Finally, the iPrOH binding site of porcine pancreatic elastase is highly homologous to human neutrophil elastase and many potent inhibitors of the later have been described, but all of them form a covalent bond with Ser195. On the basis of our calculations, we predict that a ligand that does not exploit this mechanism could achieve ΔG_{bind} up to -10.3 kcal/mol ($K_d = 34$ nM). Noncovalent inhibitors have only recently been reported and their potency is in the μM range.[36] According to our prediction, it may be possible to develop more potent noncovalent inhibitors of elastase.

Targets of Pharmacological Interest. Five proteins of pharmacological interest have been selected based on their diverse size, druggability status, and target class to further explore the reliability of our method. Solvation with an iPrOH/water mixture followed by MD simulations provided trajectories that were processed as previously described to identify hot spots and binding sites with their corresponding maximal binding affinities. Results are summarized in Table 1.

In the case of MDM2, the peptide binding groove is identified as a site capable of providing a binding free energy up to -14.6 kcal/mol. Being well below the nM range, this site can be unambiguously labeled as druggable. In agreement with this prediction, the most potent drug-like inhibitor of HDM2 (the human homologue of MDM2) has a K_d of 3 nM.[37] A second, albeit less potent, binding site (maximal $\Delta G_{bind} = -10$ kcal/

Table 1. Druggability Predictions for Five Proteins of Pharmacological Interest

target	no. atoms[a] (SASA)[b]	site	best inhibitor[c]	predicted K_d[d]	no. of sites by predicted K_d		
					≤1 nM	>1 nM ≤100 nM	>100 nM ≤10 μM
MDM2	706 (*4600*)	P53	3 nM[37]	0.02 nM	1	1	0
LFA1	1475 (*8300*)	ligand-induced	18.3 nM[39]	27 nM	0	1	0
PTP-1B	2309 (*12100*)	P-Tyr allosteric	2.2 nM[50] 8 μM[42,f]	nd[e] 0.5 μM	0	0	4
P38	2834 (*16100*)	ATP	50 pM[51]	3 pM	1	0	1
AR	2056 (*11100*)	AF2 BF3		1.0 nM 80 nM	1	1	3

[a] Number of non-hydrogen atoms. [b] Solvent-accessible surface area (Å²). [c] K_d (or K_i) of most potent small-molecule inhibitor. [d] K_d corresponding to the calculated maximal binding affinity for the specific site. [e] Not detected; note that this site is considered to be nondruggable. [f] IC$_{50}$ value (K_d not available).

mol) is found at the protein flank. This appears as a result of side chain movements of residues F86, E95, K98, and M102, which create a small cavity in the protein wall. Visualization of the different MDM2 structures deposited in the Protein Data Bank (PDB) reveal that the aforementioned residues can adopt different conformations, and one structure (PDB code 1T4F) even has a small lipophilic pocket coinciding with this putative secondary binding site. The biological relevance of this site (if any) is unknown to us.

The LFA-1/ICAM-1 complex is a second example of protein−protein interaction successfully inhibited by a drug-like compound.[31,38] In this case, inhibitors bind in a ligand-induced cavity, which is absent in the apo form of LFA-1. The flexibility of this cavity is evidenced by the fact that it closes when the water solvated protein is simulated in the absence of ligand (not shown). In the case of the binary solvent, iPrOH molecules diffuse into the active site during the equilibration protocol and keep it open for the full length of the simulation. Nevertheless, as the C-terminal α helix that forms the lid of the cavity is intrinsically flexible, the shape of the binding site changes during MD, becoming smaller than that in the crystallographic structure. Notwithstanding this limitation, the site is still predicted to be able to bind ligands with an affinity of 27 nM (−10.4 kcal/mol), which is similar to the most potent known inhibitor (K_d = 18.3 nM, as determined by isothermal titration calorimetry).[39] The protein surface does not contain other sites capable of offering tight binding.

Protein phosphatase 1B (PTP-1B) is an extremely challenging pharmacological target in spite of the fact that hundreds of inhibitors have been reported and hit rates in random screening are reasonable.[5] After years of efforts from many companies, there has been very limited success in terms of lead progression.[40] The main difficulty for this target is that binding of the inhibitors is mediated by salt−bridge interactions and inhibitors need to be constitutively charged, which greatly damages their pharmacokinetics properties.[9] Our druggability analysis reveals that there is not a single hot spot for lipophilic or neutral polar features around the phosphotyrosine binding site, which explains the total dependency of the charge to achieve potency. Interestingly, the best site, which is predicted to offer a maximal ΔG_{bind} of −9.3 kcal/mol (K_i = 180 nM), coincides with the protein−protein interface formed between PTP-1B and the insulin receptor tyrosine kinase (IRK).[41] The PTP-1B/IRK interface has a cloverleaf shape and buries a total of 1725 Å². As shown in Figure 4, the predicted binding site is located at the intersection of the three lobes, coinciding with the position of H1142 of IRK, an area delimited by V184, P187, and R268 of PTP-1B.

The second most potent binding site is located next to F196 and the C-terminal α helix, and it corresponds to a binding site of allosteric PTP-1B inhibitors.[42] The inhibitors initially reported were in the low μM range, which compares with our prediction of 500 nM, but because this is a site of considerable flexibility, it may be possible to achieve better potencies by inducing a conformational change. Finally, two weaker binding sites are detected in areas around C92 (where glycerol and 2-methyl-2.4-pentanediol have been found to bind: PDB codes 2CNG, 1SUG, and 2CM2) and near residues L234 and V249.

MAP kinase p38 has been included in this set as a representative of the kinase superfamily, which constitutes one of the main target classes, widely recognized as druggable.[43] The most potent p38 inhibitors found in the binding database[44] have a K_i of 0.05 nM (ΔG_{bind}= −14kcal/mol). This is in very good agreement with the sum of all interaction points in the active side, which add up to −15.7 kcal/mol. Most of the interaction points are located in the adenine and ribose binding pocket, which contribute −11.6 kcal/mol. The remaining ΔG_{bind} is obtained from a small lipophilic pocket formed by L74, L75, M78, and F169, which constitutes the allosteric site used by diaryl ureas.[45] The only other binding site detected on the surface of p38 is in the μM range and is located between the N-terminal tail and the C-terminal α helix.

The androgen receptor (AR) plays an essential role in prostate cancer, and antiandrogens are commonly used as a therapy.[46] As an alternative to the 5-α dihydrotestosterone (DHT) binding pocket, the activation function 2 (AF2) cleft and other regulatory surfaces involved in protein−protein binding have been proposed as targets.[47] AF2 is a binding site for coactivators, and drug-like molecules with low μM activities for the AF2 site of the thyroid hormone receptor have been discovered,[48] which sets a hopeful precedent. To investigate the potential of AF2 as a target site, we have simulated the DHT-bound ligand binding domain of AR. According to our predictions, the AF2 site is identified as capable of offering −12.3 kcal/mol for binding, which equates to a K_i of 1 nM and offers good prospects for this site. Another interesting observation is an 80 nM binding site coinciding with the recently described binding function 3 (BF3) site[49] (see Figure 5). BF3 ligands have been discovered by crystallographic fragment screening, but their potency has not been reported yet and nothing is known about its biological role. Our results suggest that this site merits further attention. Scattered over the protein surface, other sites of lower relevance have also been found. These consist of roughed surfaces or very shallow cavities located around (1) L674, T800, and I841, (2) V757, M761, Y763, and V769, (3) F878 and P904, and (4) I835,

F856, and F916. The first one contains both polar and apolar interaction points, while the other three are exclusively of an apolar nature.

Discussion

Biological macromolecules spontaneously associate with themselves or other molecules (such as metabolites or xenobiotics) because part of their surface prefers to form interactions with organic molecules rather than with water. These inbuilt solvation preferences are, therefore, the hallmark of a binding site and should be very useful to predict their existence. Here we have used MD simulations of a mixture of water and organic solvent (20% iPrOH) to elucidate solvation patterns on the surface of proteins. As a first test, we have investigated several experimentally determined binding sites of the organic molecule of choice (iPrOH). Reproducing solvent mapping experiments computationally is not an easy task, as already shown by English et al., who demonstrated that there is no correlation between observed apparent affinities and calculated interaction energies.[21] Indeed, the free energy of binding is the result of a subtle balance between many different terms and the interaction energy is only one of them. Accounting for the desolvation effect and using some clustering techniques can increase the probability of discerning true binding sites,[25] but it remained to be seen if a method based on first principles could reproduce the observed binding preferences. Our results suggest that MD provides a reasonably good ensemble of the solute−solvent configurational space, which may be useful to detect binding sites.

We have then considered that drug-like molecules will derive a large portion of their binding free energy from hydrophobic contacts and from hydrogen bonds with polar (nonionic) features, which leads us to postulate that the observed iPrOH distribution can be used to quantify the maximal binding affinity that a drug-like molecule can attain for a given site. For the proteins in our test set, this is certainly the case, as the predicted maximal ΔG_{bind} are in good agreement with the binding affinities of the most potent inhibitors. Obviously, the individual terms contributing to ΔG_{bind} of a drug-like molecule may differ significantly from those in an iPrOH free energy grid, but the results allow us to think that a similar balance is obtained. For instance, on the one hand, a drug-like ligand may suffer larger entropy costs due to loss of conformational degrees of freedom and its topology may prevent it from placing some atoms in the ideal locations, but on the other hand, it benefits from a more diverse collection of atom types which may offer better chemical complementarity to the protein than iPrOH.

A drug candidate typically binds its target with low nM potency, thus maximal affinity predictions in the sub-nM range are a good indication that the site is druggable. Nevertheless it should be noted that some drugs bind their targets with much weaker potency. Hence one should not place hard limits on the maximal binding affinity and it is important to identify all sites offering binding opportunities. This is perfectly illustrated by the PTP-1B allosteric inhibitors, which achieve biological activity in spite of their weak binding affinity.[42]

The approach described herein features several unique advantages over current methods but also some potential shortcomings that should be taken into account. Among the advantages, it is particularly noticeable that this is a nonparametric method that should be applicable to any target class, regardless of the amount of previous information available. This universal character also applies to the interaction type, as it detects hot spots for binding that may be exploited either by small molecules or by macromolecules (or both, as in the case

of MDM2). In fact, the only difference between druggable and nondruggable binding sites is that the former have a greater density of hot spots, thus resulting in more efficient binding. Another distinctive feature of our method is that binding sites appear as a result of clustering adjacent interaction points. This means that, in addition to a prediction for the whole site, one also obtains a map of the interaction preferences that may have other interesting applications. For instance, the most exothermic interaction spots could be used to define a pharmacophore or as a guide in docking.

As molecules that do not contain constitutively charged groups are more likely to be drug-like (i.e., to have adequate pharmacokinetics), we have initially ignored ionic interaction sites. Nevertheless, it may be argued that many drugs do contain ionic groups that form salt bridges or metal complexes with the protein and are crucial for binding. From a methodological viewpoint, accounting for these interaction types should be relatively straightforward, as it would only require making use of additional probe molecules. Judging from the PTP-1B example, it may seem that, at least, a certain balance between ionic and nonionic interactions may be required. Thus, from a practical point of view, even if ionic interactions are acceptable, it may be useful to know how much can be gained from lipophilic and polar neutral features.

Finally, it should be noted that as the method is based on MD, it inherits its limitations as well as its merits but, being such a widespread technique, the solutions and know-how are also at hand. For instance, the technique is computationally demanding, but there is excellent freeware to run parallel simulations. Sampling is a potentially more serious limitation, as correct predictions can only be presumed if the solvent follows a Boltzmann distribution. Both water and iPrOH molecules may be kinetically trapped in certain sites, such as buried binding sites, which would result in slow diffusion rates and insufficient sampling. We have encountered few such sites in the proteins we have studied (see Supporting Information), but enhanced sampling techniques[52] might be valuable to improve the exploration of the configurational landscape.

Conclusion

We propose a new method to detect binding sites and to quantify the maximal binding affinity that a ligand may achieve for them. In contrast with other published methods, it is based on first-principles molecular simulations and is not trained on a data set. As such, it is particularly suited to study binding sites that do not fall into the main target classes. Although it is computationally demanding, it provides a completely independent measure of druggability, which can be used instead of, or in conjunction with, higher throughput methods. Furthermore, our strategy provides very detailed information about the interaction preferences of the binding sites and can be extremely effective to give a new perspective on the target of interest.

Methods

Choice of Solvent Mixture. To map the interaction preferences of biomolecular surfaces, one can use molecular probes representing the most common chemical groups.[53] We reasoned that, to have good pharmacokinetics, a drug molecule must obtain most of its binding free energy from hydrophobic or polar neutral groups, thus a simple organic solvent could be used as the minimal expression of a drug-like molecule. As we decided to use MD for sampling, the organic molecule would also have to be small, thus ensuring a fast diffusion coefficient. Small aliphatic alcohols fulfill these conditions and have the additional advantages of being fully miscible in water and not acting as denaturants at low concentra-

tions. Our final choice of iPrOH was based on the fact that several iPrOH binding sites have been experimentally detected in proteins, which allows us to test the computational strategy. Most force-field parameters are readily available for iPrOH, as it resembles the side chain of threonine, and only the electrostatic charges had to be parametrized using the RESP procedure.[54] Table S1 of the Supporting Information lists the set of parameters used in the simulation.

A box containing 12 iPrOH molecules and 240 TIP3P water molecules, which approximately corresponds to a 20% v/v mixture, was subjected to a 5 ns MD simulation at 300 K and a constant pressure of 1 atm. Replicas of this equilibrated box were used to solvate the systems under investigation. The concentration was chosen to provide adequate sampling of the protein–isopropyl alcohol configurational space within a reasonable time scale while ensuring that the organic solvent does not disturb the dynamics of the simulated systems. Although this value has not been optimized, in all the simulations we observe a stable trajectory, with rmsd values relative to the X-ray structure similar to those typically obtained with pure aqueous solutions (Figure S1 of Supporting Information).

MD Preparation and Simulation Conditions. A representative structure of the systems under study was taken from the PDB[55] and prepared to carry out MD simulations. The preparatory steps include addition of missing atoms, removal of duplicated atoms (such in the case of double occupancies), and assignment of the most adequate protonation state of histidine residues. Crystal-lographic ions and water molecules closer than 3.5 Å from any protein atom were retained. The Leap module of the AMBER package[56] was used to include the system in a solvent-filled truncated octahedral box spanning 13 Å further from the protein. Chlorine or sodium ions are finally added to obtain an electrostati-cally neutral system. To obtain a homogeneous distribution of the solvent on the protein surface, each system was subjected to a long equilibration procedure (1.4 ns) whereby Cartesian constraints are placed on the solute atoms to prevent unfolding while the temperature is increased up to 600 K (details available in Table S3 of Supporting Information). The productive part of the simulation is carried out at constant pressure (1 atm) and temperature (300 K) for a minimum of 16 ns. All simulations are carried out under periodic boundary conditions and long-range electrostatics are accounted for using the particle-mesh Ewald summation method, as implemented in the PMEMD module of the AMBER package.[56] The coordinates of all atoms in the system were saved every 2 ps for further analysis. Calculations were performed at the Barcelona Supercomputing Centre.

MD Analysis and Hot Spot Detection. The ptraj module of the AMBER package was used to generate grids with a spacing of 0.5 Å and to count the number of times that a water oxygen, an iPrOH oxygen, or the carbon of the methyl groups of iPrOH fall in each of the grid elements (N_i). The central C atom of iPrOH is not used because it does not have a pure hydrophobic character and has a limited amount of solvent-accessible surface. To avoid grid artifacts, the population of each grid element is averaged with its neighbor positions. The expected population of each solvent type (N_o) was extracted from MD simulations of the water/iPrOH mixture, with no solute. Given N_i and N_o, eq 1 provides the free energy of binding, but it should be noted that the N_i/N_o relationship is dependent on the size of the grid elements and whether or not averaging is carried out. For this reason, it is useful to consider that, according to Kuntz & Kollman, ΔG_{bind} per atom should not be lower than -1.5 kcal/mol.[33] This upper limit (corresponding to $N_i > 12.5 N_o$) is, indeed, very rarely reached in the case of water molecules, which justifies our choice of grid parameters. Contrarily, and to our initial surprise, we found that both for Me-iPrOH and O-iPrOH the -1.5 kcal/mol threshold is surpassed far more frequently than expected. The reason for this behavior is that proteins produce a partial separation of phases, as iPrOH concen-trates around apolar surface patches and keeps away from charged areas. As a result, the reference values, which were obtained from homogeneous binary solutions, are too low and they were rescaled

in order to obtain a profile of binding free energies similar to the one for water (Figure S2 of Supporting Information).

Quantification of Maximal Affinities. Grid files containing ΔG_{bind} values of the three solvent types are read by a perl script created in our laboratory, and binding sites and their corresponding maximal binding affinities are calculated as follows:

(1) Selection of points: the grid element with the overall lowest value is selected, and all grid elements 1.4 Å around it are set to zero (this is taken as the distance of a covalent bond). The process is repeated until all points with a predicted ΔG_{bind} lower than -0.83 kcal/mol have been selected (note that this value corresponds to a population 4-fold the expected one). The points selected from the water grid are not subsequently used, but including them ensures that polar/apolar points are only selected if they have a higher affinity for the protein than water. This provides a minimum of 43 points for LFA1 and a maximum of 120 points for AR.

(2) Detection of affinity areas: A graph is created, with high affinity points considered as vertices, and nodes added between them if they are less than 2.5 Å apart. This results in a sparse graph containing small subgraphs disconnected between them. Being similar in size to a molecule used in fragment-based drug design,[57] the "affinity areas" defined by a subgraph can be thought of as fragment binding sites.

(3) Pruning: Often, an affinity area may consist of points of maximal affinity surrounded by others of lower affinity. Keeping all the points would mean that the central one losses much solvent accessibility and would result in an overestimate of ΔG_{bind}. We thus "prune" the subgraphs in a way that solvent accessibility of the highest affinity points is ensured.

(4) Clustering: In the last step, several subgraphs are merged together to form larger volumes, corresponding to the size of a typical drug-like molecule. The volume is estimated as that of a spheroid defined by the principal moments of inertia of the points being considered. For each volume, the maximal affinity is simply the sum of ΔG_{bind} of all the points included. This step can be supervised to ensure that affinity areas have been merged in a meaningful way.

Supporting Information Available: Details of the simulations are described in Tables S1 (Force-field parameters for isopropyl alcohol), S2 (PDB codes of simulated systems), and S3 (equilibra-tion protocol). An extended supplementary discussion about sampling and the use of aqueous/organic mixtures as solvent in MD simulations is also provided. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Smith, C. Drug target validation: Hitting the target. *Nature* **2003**, *422*, 341, 343, 345. passim.

(2) 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.

(3) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavail-ability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(4) Barril, X.; Soliva, R. Molecular modelling. *Mol. Biosyst.* **2006**, *2*, 660–681.

(5) 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.

(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. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(7) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.

(8) 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.

(9) Abad-Zapatero, C. Ligand efficiency indices for effective drug discovery. *Expert Opin. Drug Discovery* **2007**, *2*, 469–488.

(10) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein−protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery* **2004**, *3*, 301–317.

(11) Brown, S. P.; Hajduk, P. J. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* **2006**, *1*, 70–72.

(12) Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25–26.

(13) Maggiora, G. M. On outliers and activity cliffs—why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(14) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots—a review of the protein−protein interface determinant amino acid residues. *Proteins* **2007**, *68*, 803–812.

(15) Liepinsh, E.; Otting, G. Organic solvents identify specific ligand binding sites on protein surfaces. *Nat. Biotechnol.* **1997**, *15*, 264–268.

(16) Byerly, D. W.; McElroy, C. A.; Foster, M. P. Mapping the surface of *Escherichia coli* peptide deformylase by NMR with organic solvents. *Protein Sci.* **2002**, *11*, 1850–1853.

(17) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* **1996**, *100*, 2605–2611.

(18) Mattos, C.; Ringe, D. Locating and characterizing binding sites on proteins. *Nat. Biotechnol.* **1996**, *14*, 595–599.

(19) Ringe, D.; Mattos, C. Analysis of the binding surfaces of proteins. *Med. Res. Rev.* **1999**, *19*, 321–331.

(20) Mattos, C.; Ringe, D. Proteins in organic solvents. *Curr. Opin. Struct. Biol.* **2001**, *11*, 761–764.

(21) English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng.* **2001**, *14*, 47–59.

(22) Ho, W. C.; Luo, C.; Zhao, K.; Chai, X.; Fitzgerald, M. X.; Marmorstein, R. High-resolution structure of the p53 core domain: implications for binding small-molecule stabilizing compounds. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 1484–1493.

(23) Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *J. Mol. Biol.* **2006**, *357*, 1471–1482.

(24) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29–34.

(25) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290–4295.

(26) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. Grand canonical Monte Carlo simulation of ligand−protein binding. *J. Chem. Inf. Model.* **2006**, *46*, 231–242.

(27) Landon, M. R.; Lancia, D. R., Jr.; Yu, J.; Thiel, S. C.; Vajda, S. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J. Med. Chem.* **2007**, *50*, 1231–1240.

(28) Joseph-McCarthy, D.; Tsang, S. K.; Filman, D. J.; Hogle, J. M.; Karplus, M. Use of MCSS to design small targeted libraries: application to picornavirus ligands. *J. Am. Chem. Soc.* **2001**, *123*, 12758–12769.

(29) Chipot, C.; Pohorille, A. In *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer Series in Chemical Physics; Castleman, A. W. J., Toennies, J. P., Yamanouchi, K., Zinth, W. , Eds.; Springer: Berlin, 2007; Vol. 86, pp 517.

(30) Chene, P. Inhibition of the p53−MDM2 interaction: targeting a protein−protein interface. *Mol. Cancer. Res.* **2004**, *2*, 20–28.

(31) Pagliaro, L.; Felding, J.; Audouze, K.; Nielsen, S. J.; Terry, R. B.; Krog-Jensen, C.; Butcher, S. Emerging classes of protein−protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* **2004**, *8*, 442–449.

(32) Fasan, R.; Dias, R. L.; Moehle, K.; Zerbe, O.; Obrecht, D.; Mittl, P. R.; Grutter, M. G.; Robinson, J. A. Structure−activity studies in a family of beta-hairpin protein epitope mimetic inhibitors of the p53−HDM2 protein−protein interaction. *ChemBioChem* **2006**, *7*, 515–526.

(33) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997–10002.

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

(35) Selkti, M.; Tomas, A.; Gaucher, J. F.; Prange, T.; Fournie-Zaluski, M. C.; Chen, H.; Roques, B. P. Interactions of a new alpha-aminophosphinic derivative inside the active site of TLN (thermolysin): a model for zinc-metalloendopeptidase inhibition. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 1200–1205.

(36) Wei, L.; Gan, X.; Zhong, J.; Alliston, K. R.; Groutas, W. C. Noncovalent inhibitors of human leukocyte elastase based on the 4-imidazolidinone scaffold. *Bioorg. Med. Chem.* **2003**, *11*, 5149–5153.

(37) Ding, K.; Lu, Y.; Nikolovska-Coleska, Z.; Wang, G.; Qiu, S.; Shangary, S.; Gao, W.; Qin, D.; Stuckey, J.; Krajewski, K.; Roller, P. P.; Wang, S. Structure-based design of spiro-oxindoles as potent, specific small-molecule inhibitors of the MDM2−p53 interaction. *J. Med. Chem.* **2006**, *49*, 3432–3435.

(38) Giblin, P. A.; Lemieux, R. M. LFA-1 as a key regulator of immune function: approaches toward the development of LFA-1-based therapeutics. *Curr. Pharm. Des.* **2006**, *12*, 2771–2795.

(39) Crump, M. P.; Ceska, T. A.; Spyracopoulos, L.; Henry, A.; Archibald, S. C.; Alexander, R.; Taylor, R. J.; Findlow, S. C.; O'Connell, J.; Robinson, M. K.; Shock, A. Structure of an allosteric inhibitor of LFA-1 bound to the I-domain studied by crystallography, NMR, and calorimetry. *Biochemistry* **2004**, *43*, 2394–2404.

(40) Kasibhatla, B.; Wos, J.; Peters, K. G. Targeting protein tyrosine phosphatase to enhance insulin action for the potential treatment of diabetes. *Curr. Opin. Invest. Drugs* **2007**, *8*, 805–813.

(41) Li, S.; Depetris, R. S.; Barford, D.; Chernoff, J.; Hubbard, S. R. Crystal structure of a complex between protein tyrosine phosphatase 1B and the insulin receptor tyrosine kinase. *Structure* **2005**, *13*, 1643–1651.

(42) 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. *Nat. Struct. Mol. Biol* **2004**, *11*, 730–737.

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

(44) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(45) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268–272.

(46) Taplin, M. E. Drug insight: role of the androgen receptor in the development and progression of prostate cancer. *Nat. Clin. Pract. Oncol.* **2007**, *4*, 236–244.

(47) Estebanez-Perpina, E.; Jouravel, N.; Fletterick, R. J. Perspectives on designs of antiandrogens for prostate cancer. *Expert Opin. Drug Discovery* **2007**, *2*, 1341–1355.

(48) Arnold, L. A.; Estebanez-Perpina, E.; Togashi, M.; Jouravel, N.; Shelat, A.; McReynolds, A. C.; Mar, E.; Nguyen, P.; Baxter, J. D.; Fletterick, R. J.; Webb, P.; Guy, R. K. Discovery of small molecule inhibitors of the interaction of the thyroid hormone receptor with transcriptional coregulators. *J. Biol. Chem.* **2005**, *280*, 43048–43055.

(49) Estebanez-Perpina, E.; Arnold, L. A.; Nguyen, P.; Rodrigues, E. D.; Mar, E.; Bateman, R.; Pallai, P.; Shokat, K. M.; Baxter, J. D.; Guy, R. K.; Webb, P.; Fletterick, R. J. A surface on the androgen receptor that allosterically regulates coactivator binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16074–16079.

(50) Xin, Z.; Liu, G.; Abad-Zapatero, C.; Pei, Z.; Szczepankiewicz, B. G.; Li, X.; Zhang, T.; Hutchins, C. W.; Hajduk, P. J.; Ballaron, S. J.; Stashko, M. A.; Lubben, T. H.; Trevillyan, J. M.; Jirousek, M. R. Identification of a monoacid-based, cell permeable, selective inhibitor of protein tyrosine phosphatase 1B. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3947–3950.

(51) Liu, C.; Wrobleski, S. T.; Lin, J.; Ahmed, G.; Metzger, A.; Wityak, J.; Gillooly, K. M.; Shuster, D. J.; McIntyre, K. W.; Pitt, S.; Shen, D. R.; Zhang, R. F.; Zhang, H.; Doweyko, A. M.; Diller, D.; Henderson, I.; Barrish, J. C.; Dodd, J. H.; Schieven, G. L.; Leftheris, K. 5-Cyanopyrimidine derivatives as a novel class of potent, selective, and orally active inhibitors of p38alpha MAP kinase. *J. Med. Chem.* **2005**, *48*, 6261–6270.

(52) Norberg, J.; Nilsson, L. Advances in biomolecular simulations: methodology and recent applications. *Q. Rev. Biophys.* **2003**, *36*, 257–306.

(53) Sotriffer, C.; Klebe, G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* **2002**, *57*, 243–251.

(54) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(55) 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.

(56) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.

(57) Baurin, N.; Aboul-Ela, F.; Barril, X.; Davis, B.; Drysdale, M.; Dymock, B.; Finch, H.; Fromont, C.; Richardson, C.; Simmonite, H.; Hubbard, R. E. Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2157–2166.

(58) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific: San Carlos, CA, 2002.

(59) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.