# Druggability Indices for Protein Targets Derived from NMR-Based Screening Data

Philip J. Hajduk,* Jeffrey R. Huth, and Stephen W. Fesik

*Global Pharmaceutical Research and Development, Abbott Laboratories, Abbott Park, Illinois 60064*

*Received October 28, 2004*

An analysis of heteronuclear-NMR-based screening data is used to derive relationships between the ability of small molecules to bind to a protein and various parameters that describe the protein binding site. It is found that a simple model including terms for polar and apolar surface area, surface complexity, and pocket dimensions accurately predicts the experimental screening hit rates with an $R^2$ of 0.72, an adjusted $R^2$ of 0.65, and a leave-one-out $Q^2$ of 0.56. Application of the model to predict the druggability of protein targets not used in the training set correctly classified 94% of the proteins for which high-affinity, noncovalent, druglike leads have been reported. In addition to understanding the pocket characteristics that contribute to high-affinity binding, the relationships that have been defined allow for quantitative comparative analyses of protein binding sites for use in target assessment and validation, virtual ligand screening, and structure-based drug design.

## Introduction

Understanding the fundamentals of molecular recognition is important for both biological and pharmaceutical research. Many studies of proteins in complex with their natural ligands have been performed to understand the forces that govern complex formation. Characteristics such as pocket size and geometry,[1−3] surface complexity,[4] and complementarity of shape and polarity[5] have all been proposed as factors that contribute to the binding energy. Intriguingly, it has been recognized for many protein−ligand complexes that certain regions of the binding surface contribute a disproportionate amount of the binding energy.[6] This has been reported for very large interaction surfaces, such as protein−protein interactions, as well as for much smaller surfaces, such as protein−small molecule interactions. These studies suggest that there are energetic focal points, often called "hot spots," on protein surfaces that are the major contributors to the binding energy. Interactions of the ligand with additional regions of the protein surface serve primarily to increase specificity. From a pharmaceutical perspective, it has recently been postulated that targeting these hot spots on protein surfaces with smaller (molecular weight less than 300) and more soluble (ClogP between 1 and 3) lead molecules may be a superior route to the development of therapeutic agents compared with a strategy that begins with larger, more hydrophobic leads that tend to come from high-throughput screening of random corporate repositories.[7−9] It is therefore of great interest to develop methods not only to rapidly identify the location of hot spots on protein surfaces but also to assess their capacity to efficiently bind to small organic molecules.

Over the past decade, we have applied heteronuclear-NMR-based screening against dozens of protein targets.[10−12] Our screening library consists of approximately 10 000 diverse compounds that conform to "fragmentlike" or "leadlike" characteristics, with an average molecular weight of 220 and an average ClogP of 1.5. One of the strengths of heteronuclear-NMR-based screening is that, since the entirety of the protein is monitored for perturbations upon addition of the test compound, binding to any region of the surface can in principle be detected. As a result, ligands can be characterized not only by their affinity, but also by the site to which they bind. Another advantage of NMR-based screening is that even ligands with very weak affinities can be detected. In our typical screen (utilizing test compound concentrations of 0.5−1.0 mM), ligands with $K_D$ values as high as 5 mM can be identified. All of these advantages make heteronuclear NMR an ideal tool for the identification and characterization of hot spots on protein surfaces. Here we present an analysis of NMR-based screening data against 23 different proteins in order to derive relationships between the ability of small molecules to bind to a protein and various parameters that describe the protein binding site. The resulting computational algorithm allows for a quantitative assessment of the capacity of a given binding site to bind to small, leadlike compounds with high affinity and specificity.

## Results and Discussion

**NMR-Based Identification of Hot Spots on Protein Surfaces.** In Table 1, the number of hits, binding affinities, hit rates, and binding site locations for 28 binding sites on 23 different proteins is given. The targets are from diverse protein families and include enzymes and proteins that bind to DNA, proteins, and other endogenous ligands. The hit rates for these proteins span almost 2 orders of magnitude, from less than 0.01% to 0.94%, and the observed $K_D$ values for the hits range from 10 to 5000 $\mu$M. What is striking about these data is that, across all targets, nearly 90% of the ligands identified in the screen bind to a site on

* To whom correspondence should be addressed: Abbott Laboratories, 100 Abbott Park Rd., R46Y, AP10, Abbott Park, IL 60064-6098. Phone: (847) 937-0368. Fax: (847) 938-2478. E-mail: philip.hajduk@abbott.com.

10.1021/jm049131r CCC: $30.25 
Published on Web 02/24/2005

**Table 1.** Targets, Binding Sites, and Hit Rate Data Derived from Heteronuclear-NMR-Based Screening against 23 Protein Targets

| protein no. | pocket no. | target[a] | binding site[b] | no. tested[c] | no. hits[d] | no. series[e] | hit rate | $K_D$ range ($\mu$M)[f] | high-affinity ligand?[g] | log(HR) expt[i] | log(HR) pred[j] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | AK | adenosine | 4600 | 10 | 9 | 0.22 | 80−5000 | yes | −0.66 | −0.42 |
| 2 | 2 | Akt-PH | IP3 | 8090 | 1 | 1 | 0.01 | − | | −1.91 | −1.98 |
| 3 | 3 | Bcl-xL | Bak | 9373 | 73 | 59 | 0.78 | 10−5000 | yes | −0.11 | −0.64 |
| 4 | 4 | bir3 | peptide | 8640 | 8 | 8 | 0.09 | 600−260 0 | | −1.03 | −0.72 |
| 5 | 5 | CMPK | CMP | 8090 | 6 | 3 | 0.07 | 30−240 | | −1.13 | −0.81 |
| 5 | 6 | CMPK | other[h] | 8090 | 4 | 3 | 0.05 | 30−440 | | −1.31 | −1.27 |
| 6 | 7 | E2-31 | DNA | 1532 | 3 | 3 | 0.20 | 1000−4200 | | −0.71 | −0.72 |
| 6 | 8 | E2-31 | other[h] | 1532 | 3 | 3 | 0.20 | 30−2300 | | −0.71 | −0.42 |
| 7 | 9 | ErmAM | SAH | 7233 | 7 | 7 | 0.10 | 50−3800 | | −1.01 | −0.87 |
| 8 | 10 | FBP | DNA | 8090 | 2 | 2 | 0.02 | 200−1700 | | −1.61 | −1.04 |
| 9 | 11 | FKBP | FK506 | 6950 | 65 | 60 | 0.94 | 10−5000 | yes | −0.03 | −0.24 |
| 9 | 12 | FKBP | other[h] | 6950 | 4 | 1 | 0.06 | 100−2100 | | −1.24 | −1.22 |
| 10 | 13 | HI-0065 | ADP | 8640 | 13 | 10 | 0.15 | 10−2500 | | −0.82 | −1.28 |
| 11 | 14 | LCK | pTyr | 6953 | 43 | 38 | 0.62 | 200−5000 | yes | −0.21 | −1.07 |
| 12 | 15 | LFA | IDAS | 11029 | 44 | 23 | 0.40 | 10−1000 | yes | −0.40 | −0.35 |
| 13 | 16 | MDM2 | p53 | 8640 | 28 | 14 | 0.32 | 10−420 | yes | −0.49 | −0.35 |
| 14 | 17 | MurA | UDPNAG | 9600 | 4 | 2 | 0.04 | 30−600 | | −1.38 | −1.44 |
| 15 | 18 | MurI | Glu | 8640 | 1 | 1 | 0.01 | 2000 | | −1.93 | −2.00 |
| 16 | 19 | PAK4 | ATP | 11450 | 19 | 17 | 0.17 | 20−1000 | | −0.78 | −0.63 |
| 17 | 20 | Pin1 | peptide | 7842 | 9 | 9 | 0.11 | 50−1900 | | −0.94 | −1.49 |
| 18 | 21 | PSD95 | peptide | 11759 | 0 | 0 | 0.00 | − | | −2.00 | −1.99 |
| 19 | 22 | PTP1B | catalytic pTyr | 11892 | 25 | 20 | 0.21 | 50−5000 | yes | −0.68 | −1.15 |
| 19 | 23 | PTP1B | noncatalytic pTyr | 11892 | 2 | 2 | 0.02 | 1000−5000 | | −1.77 | −1.66 |
| 20 | 24 | SARS | RNA | 8440 | 1 | 1 | 0.01 | 1000 | | −1.93 | −1.92 |
| 21 | 25 | SCD | substrate | 622 | 5 | 5 | 0.80 | 20−5000 | yes | −0.09 | −0.55 |
| 22 | 26 | survivin | Bir3 | 9370 | 1 | 1 | 0.01 | 130 | | −1.97 | −1.99 |
| 22 | 27 | survivin | other[h] | 9370 | 33 | 30 | 0.35 | 10−5000 | yes | −0.45 | −0.35 |
| 23 | 28 | UK | peptide | 1252 | 5 | 5 | 0.40 | 10−240 | yes | −0.40 | −0.81 |
| total hits at known sites: | | | | 375 | | | | | | | |
| total hits at all sites: | | | | 419 | | | | | | | |
| percent of all hits at known sites: | | | | 89.5 | | | | | | | |

[a] Target used in the NMR-based screen. See Table S1 for more information. [b] Binding site as determined by chemical shift perturbation analysis. [c] Total number of compounds used in the screen. [d] Total number of compounds that exhibited $K_D$ values less than 5 mM. [e] Number of unique chemical series represented by the hits after clustering the compounds using a Tanimoto similarity criterion of 0.75. [f] Range of $K_D$ values observed for the screening hits. [g] Denotes whether a high-affinity ($K_D$ < 300 nM), non-peptide, noncovalent inhibitor of this target is known. See Table S1 in the Supporting Information for a detailed listing. [h] Binding to a previously unidentified ligand-binding site was detected. [i] The base-10 logarithm of the experimental hit rate. A value of −2.00 was used for hit rates of 0.00%. [j] The base-10 logarithm of the predicted hit rate as described in the text. A value of −2.00 was used for predicted scores <2.00.

the protein that is known to bind to small molecules, regardless of binding affinity. These results demonstrate the exquisite selectivity of protein surfaces to bind to ligands only at very specific locations. Such observations have been previously reported in solvent mapping of proteins by NMR[13] and X-ray crystallography,[14,15] but the size and diversity of both the compound and target sets presented here indicate that this is a general phenomenon of molecular recognition that is independent of target or compound type.

It is significant to note that a high correlation is observed between the experimental NMR hit rate and the ability to identify high-affinity ($K_D$ < 300 nM), non-peptide, noncovalent inhibitors of these targets. This is shown in Table 1, where 10 of the 14 targets (71%) possessing pockets exhibiting experimental NMR hit rates greater than 0.10% ultimately yielded high-affinity drug leads, whereas no such compounds have been reported for the nine targets lacking such pockets. These data suggests that an NMR screen of a fragment library can be used as a reliable indicator of the "druggability" of a given protein target before investing in the development of complex biochemical assays or, in the case of many genomics targets, before the function of the protein is even known.

It is also instructive to consider the four binding sites identified by NMR that were previously unknown to bind to small molecules or other endogenous ligands (gray boxes in Table 1). First, it should be noted that the ligands identified for these sites did not contain unusual functional groups (data not shown) but were comprised primarily of groups that tend to bind to protein surfaces.[16] For FKBP, the second site was proximal to the known FK506 binding site and has in fact been utilized to increase binding affinity.[10] For the other three proteins (E2-31, survivin, and CMPK), the precise locations of these additional binding sites on the proteins have been determined by high-resolution NMR structural studies, and they are spatially distinct from the known ligand-binding sites (See Figure 1). Given our current findings that small organic compounds have a strong preference to bind only to energetic hot spots on protein surfaces, it is very possible that these ancillary sites play some role in the physiological function of the protein. For example, small biaryl acids have been described that bind to the region of the DNA binding domain of E2-31 that interacts with DNA (green surface in Figure 1A).[17] Intriguingly, it has been shown that this domain of E2 can interact simultaneously with DNA and with other proteins, including p/CAF[18] and PARP,[19] thereby stimulating E2-dependent transcription. Similarly, spatially disparate regions of survivin have been shown to interact with a variety of proteins, including HSP-90 (with residues Lys79−Lys90 of survivin),[20] aurora-B kinase (which phosphorylates survivin at Thr117),[21] and INCENP.[21] Thus, the multiple hot

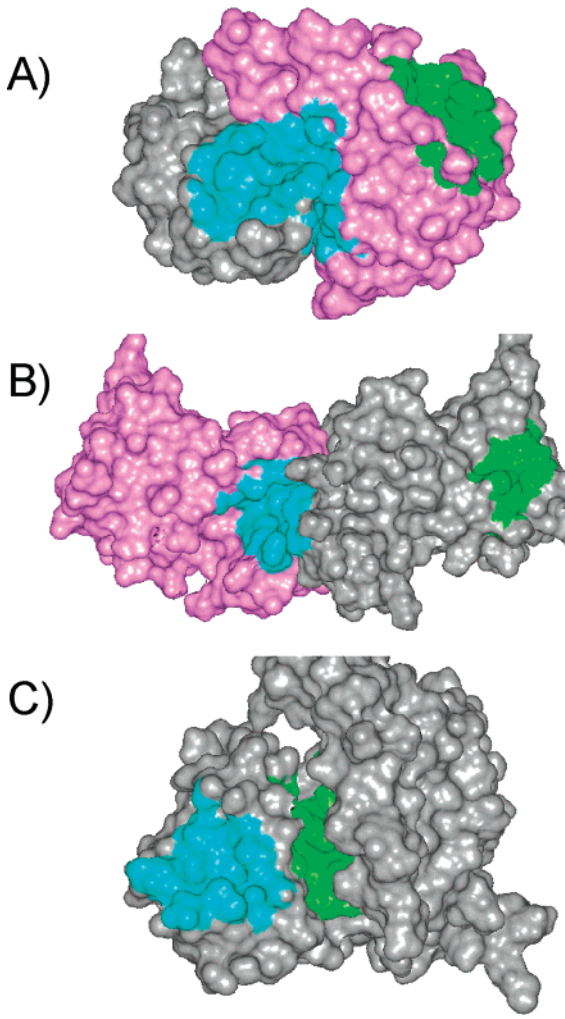


**Figure 1.** Known (green surface) and previously unknown (cyan surface) small-molecule binding sites on (A) the DNA-binding domain of E2, (B) survivin, and (C) CMPK. For E2 and survivin, which exist as dimers, one monomer is gray, while the other is pink. The green surfaces correspond to the DNA-binding site, the Bir-3 homology region, and the active site of E2, survivin, and CMPK, respectively.

spots discovered via NMR-based screening for these proteins may in fact play a role in modulating their biological function, making them suitable targets for therapeutic intervention.

**Predicting Protein Druggability.** As an alternative to executing an NMR-based screen against every potential protein target, the ability to predict with high confidence the probability that high-affinity, druglike leads can be identified for a particular target would be of tremendous value. Unfortunately, while the location of many protein binding sites can be elucidated using comparative sequence analyses, virtual docking studies,[22] or simple geometric factors,[23,24] much less is known about what influences the proclivity of a given hot spot to bind to small organic molecules. To this end, we have performed an analysis of the protein pockets and NMR screening data to try and understand the factors that influence the observed hit rate, which can be viewed as a measure of binding proclivity. For each of the protein targets, all potential pockets were identified using the flood-fill algorithm available within Insight (see Methods). This resulted in 57 potential binding sites for the 23 protein targets, including the

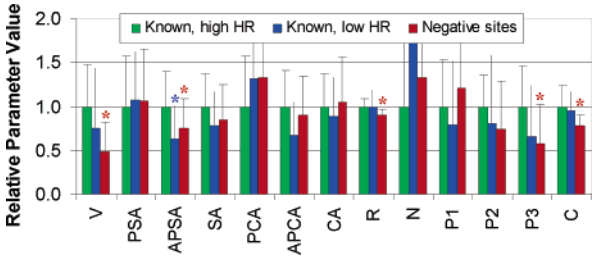


**Figure 2.** Comparison of the pocket parameter values derived from the 28 positive (known binding sites, see Table 1) and 29 negative binding sites used in the analysis. Shown in green are normalized average values for the 15 binding sites whose experimental hit rates are greater than 0.1% (see Table 1), shown in blue are values for the 13 known ligand-binding pockets whose experimental hit rates are less than 0.1% (see Table 1), and shown in red are values for the 29 negative binding sites identified from the pocket identification algorithm that are known not to bind to small molecules. All values were normalized to the average value for the 15 binding sites whose experimental hit rates are greater than 0.1% (green bars). Standard deviations in each parameter whose values are statistically different ($p < 0.05$) from the pockets with high hit rates (green bars) are denoted with an asterisk, colored by the legend for clarity. Parameters shown are volume (V), polar surface area (PSA), apolar surface area (APSA), total surface area (SA), polar contact area (PCA), apolar contact area (APCA), total contact area (CA), roughness (R), total number of charged residues (N), first (P1), second (P2) and third (P3) principal moments, and pocket compactness (C), as defined in the text.

28 known or NMR-identified ligand-binding sites shown in Table 1 (hereafter referred to as "positive" pockets) plus an additional 29 sites that were identified in the analysis but for which no compound binding could be observed in the NMR screen (hereafter referred to as "negative" pockets). Hit rates were assigned to each pocket accordingly (Table 1), with the negative pockets assigned a hit rate of 0.00%. Next, geometric parameters were extracted for each pocket. For the protein surface, parameters such as polar and apolar molecular surface area,[25] polar and apolar contact area,[25] surface roughness,[4] and the number of charged residues were calculated. For the pocket image produced within Insight, principal moments were calculated to capture the shape of the pocket, along with a parameter called pocket compactness, which was defined as the ratio of the pocket volume to pocket surface area.

A comparison of the derived parameters for the positive and negative pockets is shown in Figure 2. For ease of comparison, all of the parameters were normalized to the average value for the 15 positive pockets that exhibited high NMR hit rates (≥0.1%, see Table 1). The only statistically significant difference between positive pockets with high and low NMR hit rates was for apolar surface area (APSA, which was ~35% lower for pockets with low hit rates), although trends were observed for other parameters. For the negative pockets, multiple differences were observed, including significantly lower volume (V), apolar surface area (APSA), and roughness (R). These observations are completely consistent with the observation that endogenous ligand-binding sites tend to be the largest,[1] most hydrophobic,[22,26] and most geometrically complex[4] pockets on the protein surface. Interestingly, the negative binding sites also tended to be longer and narrower, as inferred from a significantly smaller third principal moment (P3) and compactness

**Table 2.** Regression Coefficients for the Calculation of Druggability Indices on the Basis of Pocket Parameters

| parameter | log coefficient[a] | linear coefficient[b] |
|---|---|---|
| volume | –[c] | – |
| total surface area | 2.78 (0.12) | – |
| polar surface area | – | – |
| apolar surface area | – | – |
| total contact area | – | – |
| polar contact area | −0.44 (0.38) | – |
| apolar contact area | 2.98 (0.18) | −0.023 (0.008) |
| roughness | – | 0.71 (0.11) |
| no. charged residues | – | −0.16 (0.15) |
| first principal moment | −1.03 (0.66) | – |
| second principal moment | – | – |
| third principal moment | 1.2 (1.0) | – |
| pocket compactness | 13.6 (1.5) | −14.0 (1.4) |
| constant | – | −1.11 (0.61) |

[a] Regression coefficient for the logarithm (base 10) of the corresponding parameter. 95% confidence limits are given in parentheses (See Methods). [b] Regression coefficient for the corresponding parameter. [c] No significant dependence was found (value set to zero).

(C) value. However, despite these general trends, none of these individual parameters correlated well with the experimentally determined hit rates. Thus, a regression analysis was performed using both a linear and logarithmic dependence on each of the parameters in order to identify relationships that could quantitatively correlate with the observed hit rate (see Methods). A regression analysis including eight of the 13 molecular parameters (see Table 2) yielded a good correlation with the logarithm of the observed hit rate (HR) with an $R^2$ of 0.72, an adjusted $R^2$ of 0.65, and a leave-one-out (LOO) $Q^2$ of 0.59 (see Figure 3A). If the 15 pockets that exhibit NMR hit rates >0.1% (see Table 1) are treated as true positive results, while the remaining 42 sites are treated as true negative results, then ROC curves can be generated to assess both the specificity and sensitivity of the predictions. The results of this analysis are shown in Figure 3B, where is can be observed that 67% of the true positives can be predicted before a true negative result is obtained. This is far superior to the predictive ability of the individual parameters, as is illustrated for pocket volume, apolar surface area, and roughness, where only modest enhancements in the true positive rates are observed.

**Target Characterization: Identifying Druggable Binding Sites.** From an analysis of the data (Table 1 and Figure 3), it can be observed that eight of the 10 targets for which high-affinity, druglike leads could be identified had predicted log(hit rate) values (hereafter referred to as the predicted "score") greater than −1.0, while the remaining two targets had values between −1.0 and −1.5. Thus, a protein pocket can be considered to have a "high" druggability index if the predicted score is greater than −1.0, while a "low" druggability index would be assigned for values less than −1.5. Protein pockets with intermediate values can be assigned a "moderate" druggability index. Thus, the predicted score can potentially be used in comparative analyses of all pockets on a protein surface in order to identify the sites most likely to interact with small organic compounds. To test the utility of such an approach, we identified 35 proteins not used in the training set for which high-affinity, druglike molecules have been reported and high-resolution crystal structures have been solved in

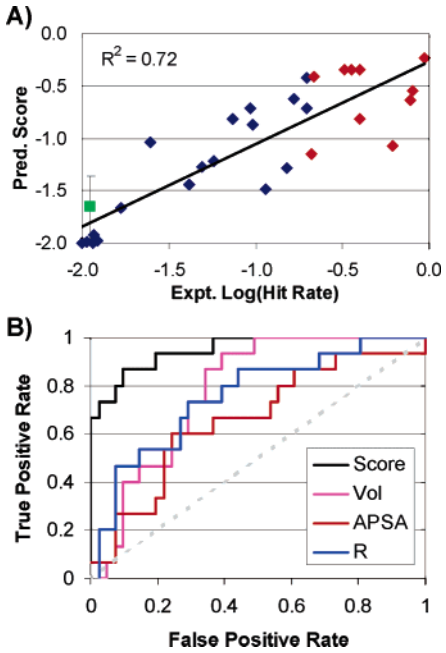


**Figure 3.** (A) Correlation between the experimental NMR hit rate and the predicted hit rate for 57 pockets on 23 protein targets as described in the text. The correlation has an $R^2$ of 0.72, an adjusted $R^2$ of 0.65, a leave-one-out cross-validated $Q^2$ of 0.56, and a standard error of 0.31 log units. The filled diamonds correspond to the 28 ligand-binding sites shown in Table 1. Solid red diamonds indicate those pockets for which high affinity ($K_D$ < 300 nM), non-peptide, noncovalent inhibitors have been reported or identified internally. For clarity, the average predicted score for the 29 protein pockets known not to bind to small organic compounds based on NMR-screening data is shown as the green square (with standard deviation). (B) ROC curves depicting the true positive vs false positive rate for pocket identification using several descriptors. In this case, true positive results were defined as the 15 pockets for which experimental hit rates of >0.1% were observed (see Table 1), while true negative results were the remaining 42 pockets used in the training set. Curves are shown for the predicted druggability score (score, black line), the pocket volume (Vol, magenta line), the apolar surface area (APSA, red line), and the pocket roughness (R, blue line).

complex with either an endogenous (e.g., ATP) or exogenous ligand (see Supporting Information, Table S2). All possible binding sites on these proteins were identified within Insight, resulting in a total of 219 binding sites for the 35 targets. Druggability scores were then calculated for each site and the known ligand-binding site was compared to the other pockets on the protein surface. As shown in Figure 4A, the known ligand-binding site was predicted to be the most druggable pocket on the surface for 25 (71%) of the proteins (green diamonds in Figure 4A). The known ligand-binding sites for eight additional proteins (blue diamonds in Figure 4A) were also predicted to be highly (score > −1.0) or moderately (score between −1.5 and −1.0) druggable, even though at least one other pocket on the surface received a higher score. Thus, a total of 33 (94%) of the 35 known ligand-binding sites were predicted to be highly or moderately druggable. ROC curves can again be generated for these data, treating the 35 known ligand-binding sites as true positive results and the remaining 184 sites as true negative results. As shown in Figure 4B, significant enhancements in the true positive rate vs the false positive rate

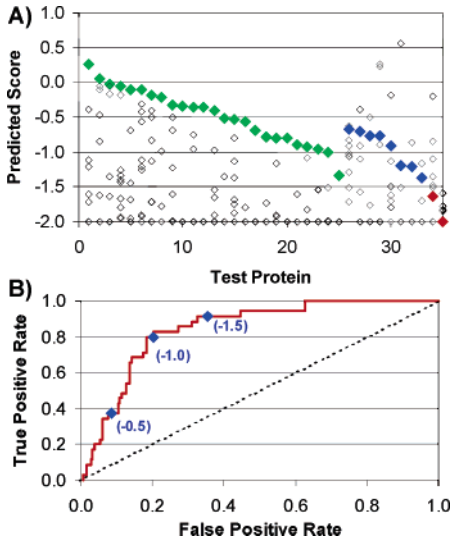


**Figure 4.** Ability of the described computational algorithm to discriminate between druggable and nondruggable binding sites on 35 proteins not used in the training set. (A) Predicted scores for the 219 pockets identified on the surfaces of the 35 targets with known high-affinity ligands. The known ligand-binding sites are denoted with a solid diamonds, while all other binding sites are shown as open diamonds. The known ligand-binding sites are ranked and colored according to whether the known site had a predicted score > −1.5 (highly to moderately druggable) and was the highest scoring pocket on the protein (green diamonds), had a predicted score > −1.5 but was not the highest scoring pocket on the protein (blue diamonds), or had a predicted score < −1.5 (red diamonds). (B) ROC curve plotting the false positive rate vs the true positive rate as a function of the score used for differentiating druggable vs nondruggable binding sites for the comparative analysis of the 219 binding sites derived from the 35 targets with known high-affinity ligands. Shown in parentheses are representative values for the cutoff score. For this analysis, true positive results were defined as the 35 known ligand-binding sites, while true negative results were the remaining 184 sites.

can be observed, providing encouragement that such calculations accurately capture many elements required for high-affinity binding to small, druglike molecules.

**Target Identification: Finding Druggable Proteins.** In addition to discriminating between multiple binding sites on a single protein, the druggability scores can also be used in screening large numbers of protein structures in an attempt to identify those most likely to be targeted with small molecule therapeutics. To assess this, an additional 37 proteins were identified for which no high-affinity, druglike ligands have been reported but for which a crystal structure in complex with either an endogenous or exogenous ligand was known (see Supporting Information, Table S2). The results of a comparative analysis of the known ligand-binding sites for the entire set of 72 proteins in the test set are shown in Figure 5. For the 35 targets with known high-affinity ligands (blue bars in Figure 5), 77% are predicted to be highly druggable, while only 6% are predicted to be not druggable. In contrast, nearly 41% of the proteins with no known high-affinity ligands (red bars in Figure 5) were assigned an index of "low." It should be noted that for the 37 targets for which no high-affinity, druglike lead could be found, we could not differentiate between those for which a search for drug leads was attempted and failed and those for which such a screen has not been performed. Thus, it is intriguing

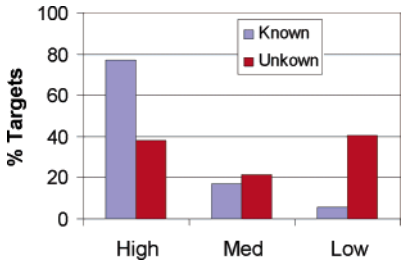


**Figure 5.** Percentage of proteins in the test set predicted to have high, moderate, and low druggability indices for those targets known (blue bars) or unknown (red bars) to have high-affinity druglike leads. Proteins were assigned a druggability index of high or low if the predicted score was greater than −1.0 or less than −1.5, respectively.

that ∼37% of these targets are predicted to be highly druggable. In any case, these results suggest that the use of druggability indices in target screening exercises can capture a high percentage of those targets most likely to bind with high affinity to small druglike compounds.

**Insights into Molecular Recognition.** As described above, many parameters and approaches have been identified that can predict the location of ligand-binding sites.[27,28] However, our analysis of the individual factors that contribute to predicted hit rates reveals that no single parameter consistently dominates the expression (see Figures 2 and 3). The dependence of the predicted hit rate on the molecular parameters described here is intuitively very satisfying (see Table 2 and Figure 6). The predicted hit rate increases logarithmically with total surface area and apolar contact area, while it decreases logarithmically with polar contact area (Figure 6A) and linearly with the number of charged residues (Figure 6B). Interestingly, the linear component for the apolar contact area begins to negatively affect the predicted hit rate beyond ∼75 Å$^2$. This suggests there is an optimal size and composition of the protein pocket that is best suited for interacting with small organic molecules. The hit rate also increases linearly with surface roughness, consistent with the notion that ligand-binding sites have high surface complexity.[4] Pocket shape also has a significant influence on the predicted hit rate. For pocket compactness, there is an optimal pocket volume to pocket surface area ratio of ∼0.4 (see Figure 6C). Larger values (corresponding to more spherical shapes) and smaller values (corresponding to more elongated shapes) have a decreased contribution to the predicted hit rate.

As shown in Figure 7, high predicted hit rates are the result of positive contributions from multiple pocket parameters, while low predicted hit rates tend to result from a significant imbalance of parameters. For example, FKBP, LFA, MDM-2, and SCD all have high experimental and predicted hit rates. However, the molecular parameters that contribute to the predicted hit rate vary for each protein. For FKBP, all parameters except for the total surface area (SA) and polar contact area (PCA) make small but positive contributions to the predicted hit rate. In contrast, the high hit rate for LFA is dominated by total surface area and pocket shape (compactness and principal moments). For MDM-2 and SCD, the lack of charged residues (Ntot) in the pocket is a significant contributor to the predicted hit rate. FBP

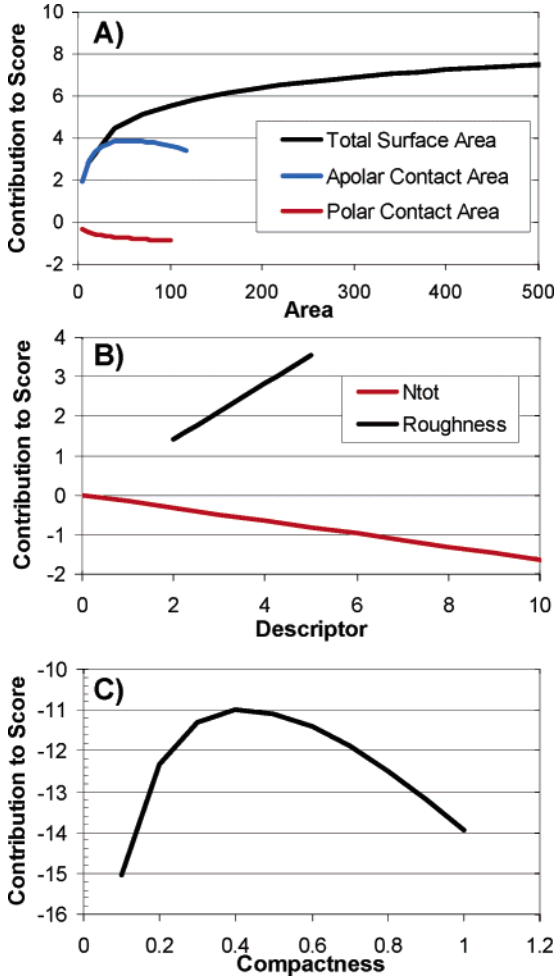


**Figure 6.** Graphical representation of the contribution of various pocket parameters to the predicted hit rate. (A) Contributions from the total molecular surface area (black line), apolar contact area (blue line), and polar contact area (red line). (B) Contributions from pocket roughness (black line) and total number of charged residues (red line). (C) Contribution from pocket compactness, defined as the ratio of the pocket volume to the pocket surface area.

is predicted to have only a moderate hit rate (log(hit rate) = −1.04). While most of the parameters have a small but positive contribution to the predicted hit rate (Figure 7), the smooth binding surface (low roughness) offsets these gains. For proteins with low predicted hit rates, one or two parameters tend to dominate. For example, the low predicted hit rate for the PH-domain of Akt is due almost exclusively to the small protein surface area at the binding site (Figure 7). The low hit rate for PSD-95 is due primarily to a small surface area and poor pocket shape (principal moments), while the lack of apolar contact area (APCA) is a significant factor for MurI (Figure 7).

Much progress has been made in understanding the forces that stabilize protein−protein interfaces. It has long been recognized that, compared to the hydrophobic core of monomeric proteins, protein−protein interfaces are relatively polar. Accordingly, analyses of the amino acids that are conserved at protein−protein interfaces have revealed high propensities not only for apolar residues but also for charged and polar amino acids, such as arginine, lysine, and aspartic acid.[29,30] Using a simple physical model, Kortemme and Baker[5] have been able to quantify the significant role that polar interactions (e.g., hydrogen bonding) and other forces play in stabilizing protein−protein interfaces. Recently, it has been reported that protein−protein docking simulations employing appropriate terms for electrostatic and hydrogen-bonding interactions can accurately predict native protein−protein interaction sites.[31] Apparently, this critical dependence on polar and charged interactions at protein−protein interfaces does not seem to be generally necessary for protein−small molecule interactions. This can first be appreciated by the fact that the amino acids that are enriched at protein−small molecule interfaces are universally hydrophobic.[26] In addition, the analysis presented here predicts at best a marginal contribution of polar residues to the druggability of a particular protein binding site. Coefficients for both the polar contact area (PCA) and total number of charged residues (Ntot) are negative, which is offset by the extent that the total surface area (SA, which has a positive coefficient) increases upon introducing these groups (see Table 2). Instead, the predicted druggability is dominated by pocket shape and hydrophobicity. While such an analysis certainly does not exclude polar or charged interactions that are highly energetically favorable, it suggests that in general such interactions may serve primarily to impart specificity rather then potency to the binding event.

## Conclusions

In summary, we have shown using heteronuclear-NMR-based screening that small organic compounds bind almost exclusively to known ligand-binding sites on proteins, regardless of binding affinity. This has important implications for uncharacterized targets derived from genomics research, where the experimental identification of therapeutically relevant binding sites can facilitate the search for novel drug leads. The relationships derived here between hit rate and pocket parameters have significant implications for understanding the fundamental principles of molecular recognition and allow for quantitative comparative analyses of protein binding sites for use in target assessment and validation, virtual ligand screening, and structure-based drug design.

## Methods

**NMR-Based Screening.** All targets were screened as previously described[11] using either $^{15}$N- or $^{13}$CH$_3$-labeled[32] protein on Bruker DRX500 or DMX500 spectrometers. Compounds were initially screened as mixtures of 10−30 compounds at concentrations of 0.4−1.0 mM each. Screening hit rates were defined as the number of individual confirmed hits with $K_D$ values below 5 mM divided by the total number of compounds screened as mixtures.

**Pocket Identification.** Pockets were identified using the ActiveSite_Search flood-fill algorithm within the Binding_Site module in InsightII (Accelrys). This is an unbiased algorithm that searches the entire protein surface and identifies all surface cavities that meet specific geometric criteria (see below). Ligands and water atoms were removed and protons were added to all proteins using the Hydrogens command within InsightII before pockets were identified. No minimization of the protein was performed, and neutral charge states were used in all instances as pockets were identified solely via geometric criteria. A grid size (parameter Grid_Size) of 1.0 Å and a cutoff size (parameter Site_CutoffSize) of 10 grid points were used in all cases. The cutoff for the site opening

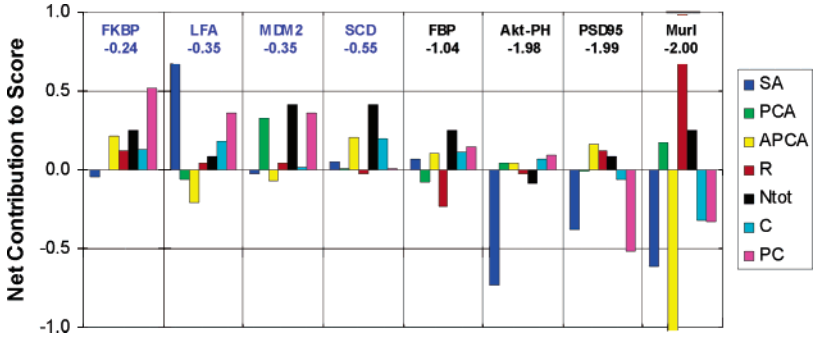


**Figure 7.** Relative contributions of seven different pocket parameters to the predicted hit rates for eight different protein targets. Shown are the contributions from total surface area (SA, blue bars), polar contact area (PCA, green bars), apolar contact area (APCA, yellow bars), roughness (R, red bars), total number of charged residues (Ntot, black bars), pocket compactness (C, cyan bars), and the first and third principal components (PC, magenta bars). Shown above each panel are the protein target and predicted score.

(parameter Site_OpenSize, which defines the maximum distance between any two exposed grid points) was varied from 3 to 7 Å. For the training set of 23 proteins, the value of the site-opening parameter that gave the maximum overlap of the pocket with the location of known ligand was used in all cases. For each protein, all additional binding sites (known experimentally not to bind to small organic molecules) were identified using the same criteria for the site opening and included in the analysis as described in the text.

**Pocket Parameters.** The volume of the pocket was taken as the number of grid points, as the grid spacing was 1.0 Å, corresponding to a volume of 1.0 Å$^3$ for each grid point. The exposed surface area for the pocket was taken as the number of sides exposed by each grid cube. Pocket compactness was then defined as the pocket volume divided by the pocket surface area. Principal moments were calculated by diagonalization of the inertia tensor, setting all weights of the grid points to 1.0.

**Protein Binding Site Parameters.** Hydrogens were removed before protein binding site parameters were calculated. The protein binding site was defined as all atoms within 4 Å of at least one pocket grid point. Total molecular and contact surface areas were calculated using the msroll command.[25] Apolar surface area was defined as that derived from carbon and sulfur atoms, while polar surface areas were defined as that derived from nitrogen and oxygen atoms. Lysine N$\zeta$ or arginine N$\eta$ atoms were counted as positive charges, while aspartic acid O$\delta$ and glutamic acid O$\epsilon$ atoms were counted as negative charges. Surface roughness was calculated as the average roughness of all atoms in the binding site using probe radii of 1.5 and 1.501 Å.[4]

**Regression Analysis.** Much of the data for the model had only a lower threshold (e.g., hit rate <0.01%, corresponding to less than 1 hit out of a 10 000 compound library) as opposed to a specific value. As standard statistical packages do not typically handle this type of data, we performed the regression analyses using the Solver command within Microsoft Excel, as has been previously described.[33] The predicted score was represented as a weighted linear combination of linear and logarithmic dependencies on each of the pocket and protein binding site parameters

$$\text{score} = \sum_{i=1}^{N} a_i X_i + b_i \log(X_i)$$

where $N$ is the number of pocket and protein binding site parameters, $X_i$ is the $i$th parameter, and $a_i$ and $b_i$ are the weighting coefficients for the linear and logarithmic terms of the $i$th parameter, respectively. So as not to penalize the scoring function for predicted hit rates less than 0.01% (score $< -2.0$), predicted scores less than this value were set to $-2.0$ before penalties were calculated. The regression analysis then minimized the $\chi^2$ between the observed and predicted hit rates. A leave-one-out (LOO) internal cross-validation of the model was performed by iteratively repeating the regression analysis while leaving out one data point. This analysis yielded a LOO $R^2$ ($Q^2$) of 0.59. Studentized and raw residual plots are shown in the Supporting Information (Figure S1). Confidence limits on the regression coefficients were estimated using constant $\chi^2$ boundaries.[34] Using this approach, each parameter was independently adjusted until the change in $\chi^2$ exceeded 18.307, which is the 95% confidence limit for a model with 10 degrees of freedom. The change in the parameter required for exceeding this value is reported as the 95% confidence limit.

**Prediction Test Set.** For the 72 proteins used for testing the predictive ability of the model, parameters for pockets and protein binding sites were obtained as described as above. Pockets that overlapped with the known ligand were saved for each value of the site opening cutoff value, which was varied from 3 to 7 Å. Scores for each of these pockets were calculated, and the pocket with the highest score was used in the analysis. For the 35 proteins with known high-affinity ligands, the value of the site-opening cutoff value that yielded the highest score for the known ligand-binding site was used to identify all binding sites on the protein surface. A maximum of 10 sites (ranked by volume) was saved for each protein.

**Acknowledgment.** The authors would like to thank Drs. Ed Olejniczak, Steve Muchmore, and Scott Brown, for helpful discussions and critical reading of the manuscript, and Dr. Yvonne Martin, for helpful discussions regarding the statistical analysis.

**Supporting Information Available:** Tables listing the proteins used in the analysis along with available information on small-molecule ligands or inhibitors and Studentized and raw residual plots from the regression analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein Clefts in Molecular Recognition and Function. *Protein Sci.* **1996**, *5*, 2438−2452.
(2) Peters, K. P.; Fauck, J.; Frommel, C. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using Only Geometric Criteria. *J. Mol. Biol.* **1996**, *256*, 201−213.
(3) Brady, G. P.; Stouten, P. F. W. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput.-Aided Mol. Design* **2000**, *14*, 383−401.
(4) Pettit, F. K.; Bowie, J. U. Protein Surface Roughness and Small Molecular Binding Sites. *J. Mol. Biol.* **1999**, *285*, 1377−1382.
(5) Kortemme, T.; Baker, D. A Simple Physical Model for Binding Energy Hot Spots in Protein−Protein Complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14116−14121.
(6) DeLano, W. L. Unraveling Hot Spots in Binding INterfaces: Progress and Challenges. *Curr. Opin. Struct. Biol.* **2002**, *12*, 14−20.
(7) Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins. *Science* **1997**, *278*, 497−499.

(8) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(9) Carr, R.; Jhoti, H. Structure-based screening of low-affinity compounds. *Drug Discovery Today* **2002**, *7*, 522−527.
(10) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.
(11) Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. NMR−Based Screening in Drug Discovery. *Quart. Rev. Biophys.* **1999**, *32*, 211−240.
(12) Huth, J. R.; Sun, C. Utility of NMR in Lead Optimization: Fragment-Based Approaches. *Comb. Chem. High Throughput Screening* **2002**, *5*, 631−644.
(13) Liepinsh, E.; Otting, G. Organic Solvents Identify Specific Ligand Binding Sites on Protein Surfaces. *Nature Biotechnol.* **1997**, *15*, 264−268.
(14) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffrey, C. J.; Mattos, C.; et al. An Experimental Approach To Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605−2611.
(15) 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.
(16) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged Molecules for Protein Binding Identified from NMR−Based Screening. *J. Med. Chem.* **2000**, *43*, 3443−3447.
(17) Hajduk, P. J.; Dinges, J.; Miknis, G. F.; Merlock, M.; Middleton, T.; et al. NMR−Based Discovery of Lead Inhibitors that Block DNA Binding of the Human Papillomavirus E2 Protein. *J. Med. Chem.* **1997**, *40*, 3144−3150.
(18) Lee, D.; Hwang, S. G.; Kim, J.; Choe, J. Functional Interaction Between p/CAF and Human Papillomavirus E2 Protein. *J. Biol. Chem.* **2002**, *277*, 6483−6489.
(19) Lee, D.; Kim, J. W.; Kim, K.; Joe, C. O.; Schreiber, V.; et al. Functional Interaction Between Human Papillomavirus Type 18 E2 and Poly(ADP-ribose) Polymerase 1. *Oncogene* **2002**, *21*, 5877−5885.
(20) Fortugno, P.; Beltrami, E.; Plescia, J.; Fontana, J.; Pradhan, D.; et al. Regulation of Survivin Function by Hsp90. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *25*, 13791−13796.
(21) Wheatley, S. P.; Henzings, A. J.; Dodson, H.; Khaled, W.; Earnshaw, W. C. Aurora-B Phosphorylation in vitro Identifies a Residue of Survivin that is Essential for its Localization and Binding to Inner Centromere Protein (INCENP) in vivo. *J. Biol. Chem.* **2004**, *279*, 5655−5660.
(22) Ruppert, J.; Welch, W.; Jain, A. N. Automatic Identification and Representation of Protein Binding Sites for Molecular Docking. *Protein Sci.* **1997**, *6*, 524−533.
(23) 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.
(24) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand Binding: Functional Site Location, Similarity, and Docking. *Curr. Opin. Struct. Biol.* **2003**, *13*, 389−395.
(25) Connolly, M. L. *msroll*; 3.8 ed.; Menlo Park, CA.
(26) Hajduk, P. J.; Mack, J. C.; Olejniczak, E. T.; Park, C.; Dandliker, P. J.; et al. SOS−NMR: A Saturation Transfer NMR−Based Method for Determining the Structures of Protein−Ligand Complexes. *J. Am. Chem. Soc.* **2004**, *126*, 2390−2398.
(27) Jones, S.; Thornton, J. M. Searching for Functional Sites in Protein Structures. *Curr. Opin. Chem. Biol.* **2004**, *8*, 3−7.
(28) Sotriffer, C.; Klebe, G. Identification and Mapping of Small-Molecule Binding Sites in Proteins: Computational Tools for Structure-Based Drug Design. *Il Farmaco* **2002**, *57*, 243−251.
(29) Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.
(30) Hu, Z.; Ma, B.; Wolfson, H.; Nussinov, R. Conservation of Polar Residues as Hot Spots at Protein Interfaces. *Proteins: Struct. Funct. Genet.* **2000**, *39*, 331−342.
(31) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. Identification of Protein−Protein Interaction Sites from Docking Energy Landscapes. *J. Mol. Biol.* **2004**, *335*, 843−865.
(32) Hajduk, P. J.; Augeri, D. A.; Mack, J.; Mendoza, R.; Yang, J.; et al. NMR−Based Screening of Proteins Containing 13C-Labeled Methyl Groups. *J. Med. Chem.* **2000**, *122*, 7898−7904.
(33) Hajduk, P. J.; Mendoza, R.; Petros, A. M.; Huth, J. R.; Bures, M.; et al. Ligand Binding To Domain-3 Of Human Serum Albumin: A Chemometric Analysis. *J. Comput.-Aided Mol. Design* **2003**, *17*, 93−102.
(34) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vettering, W. T. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: New York, 1986.

JM049131R