Drug-like Density: A Method of Quantifying the "Bindability" of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank

Robert P. Sheridan,*,† Vladimir N. Maiorov,† M. Katharine Holloway,‡ Wendy D. Cornell,† and Ying-Duo Gao†

Chemistry Modeling and Informatics Department, Merck Research Laboratories, Rahway, New Jersey 07065, United States, and Chemistry Modeling and Informatics Department, Merck Research Laboratories, West Point, Pennsylvania 19486, United States

Received August 13, 2010

One approach to estimating the "chemical tractability" of a candidate protein target where we know the atomic resolution structure is to examine the physical properties of potential binding sites. A number of other workers have addressed this issue. We characterize ~290 000 "pockets" from ~42 000 protein crystal structures in terms of a three parameter "pocket space": volume, buriedness, and hydrophobicity. A metric DLID (drug-like density) measures how likely a pocket is to bind a drug-like molecule. This is calculated from the count of other pockets in its local neighborhood in pocket space that contain drug-like cocrystallized ligands and the count of total pockets in the neighborhood. Surprisingly, despite being defined locally, a global trend in DLID can be predicted by a simple linear regression on log(volume), buriedness, and hydrophobicity. Two levels of simplification are necessary to relate the DLID of individual pockets to "targets": taking the best DLID per Protein Data Bank (PDB) entry (because any given crystal structure can have many pockets), and taking the median DLID over all PDB entries for the same target (because different crystal structures of the same protein can vary because of artifacts and real conformational changes). We can show that median DLIDs for targets that are detectably homologous in sequence are reasonably similar and that median DLIDs correlate with the "druggability" estimate of Cheng et al. (*Nature Biotechnology* **2007**, *25*, 71−75).

INTRODUCTION

Much hope centers around the idea that one can assign a reliable "druggability" measure to a gene target so that one may direct one's discovery effort to the targets most likely to result in a drug. The literature on this topic has exploded in the past 5 years.[1−22] Cheng[6] provides a very up-to-date and readable review of the issues. There are two major approaches to estimating the druggability of a gene, one based on known target/drug associations, and the other based on the structure of protein targets. In the first approach[1,3,4,7,9−11] given a gene sequence A, one looks for a homologous protein target B for which one or more drugs are already known. For example, if A is homologous to proteins B that are G-protein coupled receptors (GPCRs), for which many drugs are known, then presumably A represents a "druggable" target. The hope is that one may even try compounds that bind to B as candidates for A. In the second approach,[12−22] one looks for a homologous protein target B for which an atomic level structure is known. One inspects B for binding sites likely to bind drug-like molecules tightly. The definition of "drug-like" in this context is variable, but usually one means a molecule with molecular weight < 500, moderate lipophilicity, etc. If binding a ligand to such a site could affect the function of the protein as an inhibitor or an allosteric modulator, then A could be considered druggable. The paper of Cheng et al.[15] is one of the most cited and one of the first to address druggability using physical properties of binding sites, although the older concept of the "beautiful active site" embodies a similar idea.[23]

Both of these approaches have the same type of limitations:

(1) They are historic, in the sense that both extrapolate druggability from examples of what has been successful in the past.

(2) They make the assumption that homologous proteins behave similarly in terms of the types of molecules they bind. There is much debate as to whether this is really true.[24] There is some indication that it might be true for GPCRs.[25] In contrast, protein kinases that have very similar sequences, even in the active site, may not bind the same classes of molecules.[26]

(3) The term "druggable" implies too much. Getting to an actual drug involves many hurdles that are almost impossible to predict in advance and involve properties of small molecules as well as the target. For example, there is nothing in those approaches that determines whether the active site in the target is different enough from related targets that selectivity is possible. It is also not clear that there will be the desired therapeutic effect in vivo even if we can find drug-like molecules to bind to the target protein.

The best that can be promised by the two approaches discussed above, especially the second, is to find targets that are "bindable" by drug-like molecules. One common phrase for this is "chemically tractable". Our current effort is in the spirit of others that use physical properties of active sites

* Corresponding author phone: 732-594-3859; fax: 732-594-4224; e-mail: sheridan@merck.com.
† Merck Research Laboratories, Rahway, NJ.
‡ Merck Research Laboratories, West Point, PA.

10.1021/ci100312t
Published on Web 10/26/2010

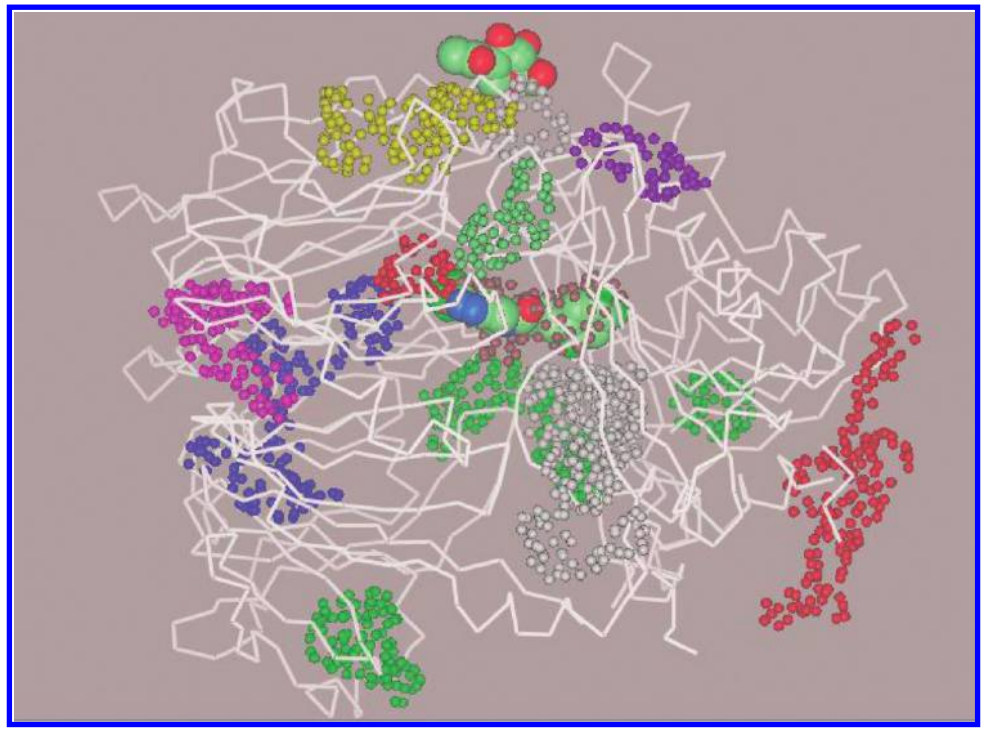

Figure 1. Pockets generated for PDB entry 1X70 (DPP4). Heterogroups associated with at least one pocket are displayed as CPK models. The group at the top is NAG, a covalently bound sugar, and the group in the center is an inhibitor.

to predict "bindability". However, whereas most previous work has used fairly small hand-curated sets of protein structures, sometimes confined to those sites that already have a bound ligand, we resolved to examine as many diverse protein structures as possible and also to include potential sites that do not already have a ligand. We thought this was necessary to compile as big a set of homologues to potential gene targets as possible.

In this paper we generate a large database of protein pockets from protein crystal structures and compile a database of ligands (some of which are drug-like) associated with those pockets. One can imagine a three-dimensional pocket space defined by volume, buriedness, and hydrophobicity and each pocket representing a location in that space. We introduce a bindability metric called "drug-like density" (DLID) that depends on the local density in pocket space of pockets containing drug-like ligands vs the density of total pockets.

METHODS

PDB Entries. The entire set of entries in the Protein Data Bank[27] (PDB) as of September 2008 was analyzed to generate a list of protein crystal structures that had ≤3 Å resolution. This list includes ∼42 000 out of ∼55 000 PDB entries.

How To Generate Pockets. Because we did not want to be limited to extracting information only from crystal structures with ligands, as in most earlier studies, we needed a way to define binding sites independent of the particular ligand in a given crystal structure, or there being any ligand present at all. There are now a number of algorithms for perceiving concavities in protein surfaces.[28−38] The one available to us is the icmPocketFinder algorithm[32] in the ICM package,[39] which generates a set of points at the surface of a virtual pocket that fills the concavity. Henceforth, we will speak of pockets instead of binding sites.

icmPocketFinder also returns the volume of the pocket and the identity of chain(s) forming the pocket. For any given PDB entry there will be many pockets in the range of 100−3000 Å^3, which is the default range for icmPocketFinder. An example of pockets generated for a PDB entry is shown in Figure 1. Details about the processing of PDB structures are in the following sections.

Protein Chains To Include in the Pocket Calculation: Biounits. Each PDB entry can contain a number of polypeptide chains. Sometimes two or more chains may form a meaningful biological unit, and all should be included. On the other hand, some chains are just duplicates in different asymmetrical units in the crystal structure. We need to know the proper biological unit when calculating pockets. For example, in HIV protease (a homodimer), the only meaningful pocket is formed by two chains A and B. To address this issue, we generally used the "biounit" information maintained by the PDB. The biounit for a PDB entry is a list of chains that the author feels are a biological or functional unit. There may be several biounit variants per PDB entry. We used the first listed, which is presumably the most reliable in the opinion of the author or the curators of PDB.

There are two situations when we used "raw" (all coordinate) PDB entries instead of a biounit. The first is a case of relatively large biounits with 3000 or more amino acid residues. The second is when a biounit contains more than one mention of the same chain. Either case interfered with our ability to process the coordinates and extract pockets via biounits.

Atoms To Include as Part of the Protein When Calculating Pockets. The PDB has no native concept of "ligand", only "atom" vs "heteroatom", i.e., peptide vs nonpeptide.

Therefore there is no direct way of distinguishing cofactors from "true" ligands such as inhibitors or allosteric modulators that are of interest as potential drugs. During preparation of a protein structure for pocket calculations, all PDB hetero compounds and biopolymer nonpolypeptide chains (e.g., DNA or RNA) were deleted. (Groups with covalent links to the protein were not treated differently.) The exception is heme or heme-like groups. This exception was made because we assume heme to almost always be a more or less permanent part of the protein. This is the list of heme-like "residues": HEM, MNR, COH, MNH, HNI, MHM, CCH, HEB, FDD, HEV, DDH, ZEM, FEC, HEA, FDE, 2FH, HE6, CLN, DEU, HIF, 1FH, VER, MP1, HME, CL1, CL2, CHL, CLA, HDM, HEC, DHE. All other PDB hetero compounds were assumed to be potential ligands for the purposes of our work. It is conceivable that one could include common cofactors other than heme as part of the protein (e.g., NAD in dihydrofolate reductase), but in practice it is very hard to automatically determine whether a particular hetero compound is a substrate/inhibitor or cofactor for a particular PDB entry (except for heme). Also the majority of PDB entries of proteins that use cofactors are not cocrystallized with the cofactor (except for heme). Given the large number of PDB entries that had to be processed, we decided that it was not practical to automatically handle any other cofactor but heme. After we had finished this work we found that Le Guilloux et al.[35] made a similar choice for similar reasons. Another reason to allow for the removal of cofactors is that many drugs bind to the cofactor site rather than the substrate site. Examples are folate mimics in thymidylate synthase and ATP-site inhibitors for protein kinases.

Similarly, it may not be easy to distinguish a peptide ligand from a part of the protein. This is especially important because where a peptide ligand is interpreted as part of a protein, the potential binding site containing the ligand is not recognized. Our rule was that short peptides ($\leq$7 residues long) are assumed to be ligands. Longer peptides are assumed to be part of the protein.

How "Shells" Are Assigned to Pockets. Protein residues with any atom within 3.5 Å of any pocket surface point are assigned to a "shell" surrounding that pocket. Any particular atom in the protein can be assigned to more than one shell.

How Ligands Are Assigned to Pockets. (1) All the HETATM records are extracted from the PDB entry and temporarily merged into a single molecule. (We assume the bond orders for the ligand in the PDB residue library are correct.)

(2) Bonds are added where there is an interresidue distance $<$ 1.6 Å for first row elements and $<$2 Å for second row elements. This rescues the cases where a single ligand molecule is made of multiple residues; otherwise they would be later incorrectly extracted as separate fragments.

(3) The molecule is divided again into individual bonded fragments.

(4) For each fragment one measures what fraction of the atoms is within 2.5 Å of any pocket surface point. If the fraction for a fragment $\geq$ 0.1, that fragment is assigned to the pocket. It is possible for a ligand to be assigned to more than one pocket if it bridges two or more concavities.

(5) All fragments assigned to a pocket are merged into one molecule for the purpose of storage with the pockets database. There is often more than one fragment per pocket.

To be as general as possible, we made no attempt to discard fragments that would normally not be considered proper ligands (e.g., sulfate, glycol, etc.) since in some cases deciding which are "proper" is difficult without specific knowledge about the protein in the crystal structure. For instance, sulfate could be a nonconsequential part of the crystallization conditions or a true ligand in case of sulfate-binding protein. It was decided that such filtering was better left to postprocessing after the database was built.

How Pocket Characteristics Are Calculated. The literature contains a variety of characteristics that can be calculated for pockets that seem to be predictive of whether that pocket binds a drug-like molecule. Characteristics that need molecular mechanics simulation[20,31] are obviously much too slow for the number of pockets we are dealing with here. Other workers have used semiempirical desolvation functions[15] or amino acid composition[16] as metrics. The pocket characteristics that are most studied[12,18,19,21,22] are volume (or area), buriedness (or exposure), hydrophobicity (or polarity), and some flavor of surface roughness or "compactness" (or sphericity). We found that three parameters, volume, buriedness, and hydrophobicity, are easy to calculate, easy to understand, and sufficient for our purposes. This is how we calculated these parameters:

(1) The pocket volume is returned by icmPocketFinder directly.

(2) The fraction buried or buriedness is calculated as follows: One measures the solvent accessible surface area of the pocket (probe radius, 1.4) in isolation. Then one measures the solvent accessible surface area of the pocket covered by its shell. The ratio of the second number to the first is the fraction buried. The lowest possible value is 0.5; i.e. the pocket is completely open and the surface flat. The highest is 1.0, i.e., completely buried.

(3) Fraction hydrophobic or hydrophobicity is the fraction of the pocket surface in contact with hydrophobic protein atoms of its shell. The definition of hydrophobic atom is given in ref 40. Hydrophobicity can vary from 0 to 1.

(4) Ligand contact ratio is the fraction of the total number of atoms in the ligand associated with the pocket (which may contain more than one fragment) that are within 2 Å of any of the pocket surface points. A ligand contact ratio of 1 means the ligand is effectively entirely within the pocket.

Pocket Database. Each pocket is assigned a name of the form {PDBcode}_OBJ_{number}. Associated with each pocket is (a) a set of coordinates representing the pocket surface, plus information about which chain(s) of the protein form the pocket and the volume of the pocket (these are directly from icmPocketFinder); (b) a "shell" of the name {PDBcode}_MOL_{number} consisting of the coordinates of protein atoms in the shell; (c) any ligands associated with the pocket written as a single molecule with the name {PDBcode}_LIG_{number}. There may be more than one "fragment" (continuously bonded set of atoms) present; for example, the largest pocket for the PDB entry 3dfr contains the fragments corresponding to NAD and methotrexate).

Property Assignment of Ligands. The following properties were calculated for each ligand fragment assumed to be at pH 7.4: (1) number of non-hydrogen atoms; (2) number of H-bond donors (each individual polar H counts as a separate donor); (3) number of H-bond acceptors; (4) sum of donors or acceptors divided by the number of non-

hydrogen atoms; (5) Klopman log P;[41] (6) number of rotatable bonds divided by the total number of nonterminal bonds.

The distribution of these properties were compared with the distribution of those properties in "oral drugs" from the PDR[42] and expressed as percentiles. For a molecule that had the number of non-hydrogens equal to the median value of non-hydrogens in oral compounds, the percentile would be 50. One can calculate the CI "centrifugal index", an empirical metric of how far a compound is from the median physical properties of oral compounds as

$$\mathrm{CI}_{(i)} = \frac{\sum_k (0.02\ \mathrm{abs}(\mathrm{percentile}(i,k) - 50))^3}{6}$$

where i is the compound and k is one of the six properties.

Defining Drug-like Ligands. There are several bodies of data one can consider for a calibration of bindability (lists of druggable targets as in Cheng et al.,[15] tables of experimental binding energies in databases such as MOAD,[43] etc.), but the one for which the most types of proteins are covered is the set of ligands that are already associated with pockets in our pocket database. A bindability metric could be defined as follows: Given a pocket of x volume, y buriedness, and z hydrophobicity, what is the probability that it could bind a drug-like ligand? In the spirit of "agnosticism", all one can do is compare the pocket at x, y, z to other pockets and say something like the following: pockets close in pocket space to x, y, z have a drug-like ligand this fraction of the time. This requires us to determine which pockets have at least one fragment that is drug-like, where a fragment is a continuously bonded set of atoms.

There are many possible definitions of drug-like. We used the following:

(1) The fragment has to be within the 5th to 95th percentile of all six computable physical properties in the preceding Property Assignment of Ligands: number of non-hydrogens, 10−43; number of donors, 0−6; number of acceptors, 0−9; Klopman log P, −2.4 to +6.7; normalized polarity, 0.1−0.6; normalized bond flexibility, 0.1−0.4.

(2) The fragment cannot be among the HETATM types that occur at least 50 times in the PDB. This eliminates common cofactors or other groups. Examples are ATP, NAD, and NAG, etc.

(3) At least 50% of the atoms of all the fragments are contained within the pocket.

Having done this, we find ∼5700 drug-like ligand containing (DLLC) pockets. These are listed in the Supporting Information.

Drug-like Density. There are a number of ways one can generate a score based on the data on DLLC pockets. Here we introduce drug-like density (DLID). Imagine each pocket in a three-dimensional pocket space defined by its volume (range, 100−3000 Å^3), buriedness (0.5−1), and hydrophobicity (0−1). One can define the "neighborhood" of the pocket as being within one "grid spacing" in that pocket space, the grid spacing being 100, 0.025, and 0.025, respectively, for those dimensions. One can count the number of DLLC pockets in the neighborhood, and the total number of pockets in the neighborhood, both numbers excluding the pocket itself. The ratio of those numbers is a measure of the relative density of DLLC pockets in the neighborhood of the pocket being examined. One can express this as log(ratio). This should be corrected by subtracting the log of the ratio of all DLLC pockets to all pockets. The corrected log(ratio) is the DLID for the pocket. A DLID of zero means that one sees the ratio of drug-like ligands in the neighborhood as expected by chance. One must introduce two arbitrary parameters. One is the minimum number of total neighbors one must have to consider the statistics "good enough"; here we used 30. Given that number, we cannot calculate the DLID for 1.6% of the pockets. The second parameter is necessary to avoid taking the log of zero when the number of DLLC pockets in a neighborhood is zero. In those cases we replace zero with 0.1.

Instead of sampling pocket space at the location of real pockets, alternatively one can sample the space uniformly as in a grid and calculate the DLID for each grid point based on the neighbors for the grid point. Similarly, one can calculate the volume, buriedness, and hydrophobicity of a protein pocket not in the original pocket database and calculate the DLID on the basis of its neighbors.

Clustering of PDB Entries into "Targets". Ultimately, applications of druggability refer to targets and not individual crystal structures. Therefore we need to group PDB entries into targets. However, "target" does not have an objective definition everyone can agree on. For example does the target "THROMBIN" include not only human thrombin, but that of other species, or include very closely related serine proteases that might not be called "thrombin"? What about mutants of thrombin? It would be impractical to hand-code target lists given the full diversity of proteins in the PDB. Fortunately, the PDB provides clusters of PDB entries based on the sequence identity for individual protein chains at a number of cutoffs of sequence identity: 100, 95, 90, 75, and 50%. We decided that clustering at 90% sequence identity was a reasonable compromise between being too restrictive and including too much and that such clusters were reasonable approximations to targets. One should note that given this definition, mutants will be included and proteins from more than a single closely related species may be included as the same target.

A PDB entry can be associated with more than one cluster if it contains more than one nonidentical chain. There can be multiple chains if there is more than one subunit (e.g., α and β chains in hemoglobin), or sometimes the second chain is merely a small peptide ligand. We defined the "official" cluster for a PDB entry to be that of its longest chain. We found that there are ∼14 000 clusters for which we can assign at least one PDB entry with a DLID. About 7900 clusters contain only a single PDB entry. The largest cluster (T4-lysozyme) contains 409 PDB entries. Assignment of PDB entries to clusters is given in the Supporting Information.

RESULTS

Universe of Pockets and Their DLIDs. We examined ∼42 000 PDB entries and produced a total of ∼290 000 pockets in the range of 100−3000 Å^3. Larger or smaller pockets were ignored. Interestingly, about 800 PDB entries turned out to have no pockets in that range. Only ∼50 000 of the pockets contain at least one ligand of any kind, and only ∼5700 pockets contain a drug-like ligand by our

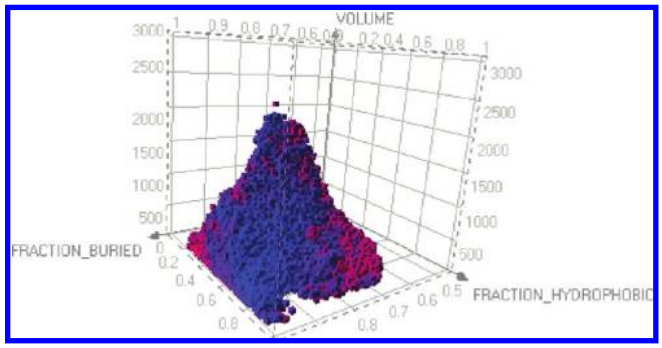

Figure 2. Three-dimensional pocket space consisting of volume, buriedness, and hydrophobicity. Each point represents a pocket. Coloring is by DLID, with blue being the most "drug-dense." Pockets with too few neighbors to calculate a DLID are omitted. Most of the red points are on the side away from the viewer.

definition. The ratio of DLLC pockets to all pockets is ~1.9%, and log(0.019) = −1.71, so one subtracts that correction to obtain the DLID for a given pocket:

$$\text{DLID} = \log(\text{DLLC neighbors/all neighbors}) + 1.71$$

The list of pockets, their properties, and DLIDs is given in the Supporting Information. Figure 2 shows a plot of all pockets in pocket space, colored by the DLID, with blue being the highest DLIDs and red being the lowest. There is clearly discrimination; some parts of space are very blue and some very red, and the blue region is more or less contiguous. No particular spot in this space is dominated by any given target.

Figure 3 is a histogram of the DLID values. For comparison is shown the distribution of DLID expected if DLLC pockets had the same distribution in pocket space as all pockets (i.e., by randomly picking ~5700 pockets from the total set of pockets and calling them DLLC pockets). The fact that the distributions are very different implies that one can predict whether a pocket binds a drug-like molecule given the pocket space. Looking at Figure 3, a reasonable cutoff for deciding whether a DLID implies that a pocket is bindable would be 0.5, which implies a 3-fold higher density than expected.

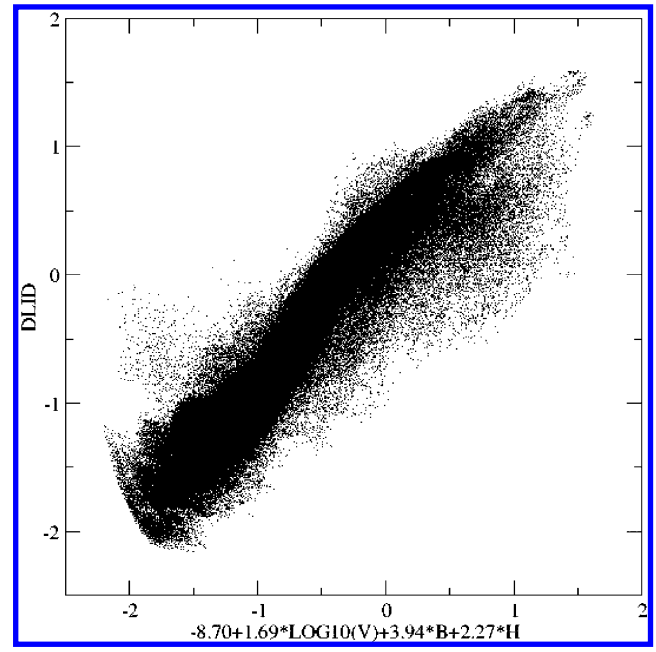

Figure 4. DLID vs the DLID calculated by a simple linear formula based on volume, buriedness, and hydrophobicity. Pockets where the number of neighbors with a drug-like ligand equals zero are omitted. All the outlier points at the lower left are cases where the number of neighbors with a drug-like ligand equals 1 or 2.

A simple equation linking pocket space with DLID is the following:

$$\text{DLID} = -8.70 + 1.71\log(\text{volume}) + 3.94(\text{buriedness}) + 2.27(\text{hydrophobicity})$$

For that equation, pockets with zero DLLC neighbors were ignored because those are given somewhat fictitious DLIDs to avoid taking the log of zero. The relationship of the real and fit DLID is shown in Figure 4. The Pearson correlation R of the fit is 0.93. The fact that DLID appears to be nearly additive as a global trend was not necessarily expected, since DLID was defined by local densities as in Figure 2. This is interesting because it implies that these parameters can compensate for each other. For example, a low volume

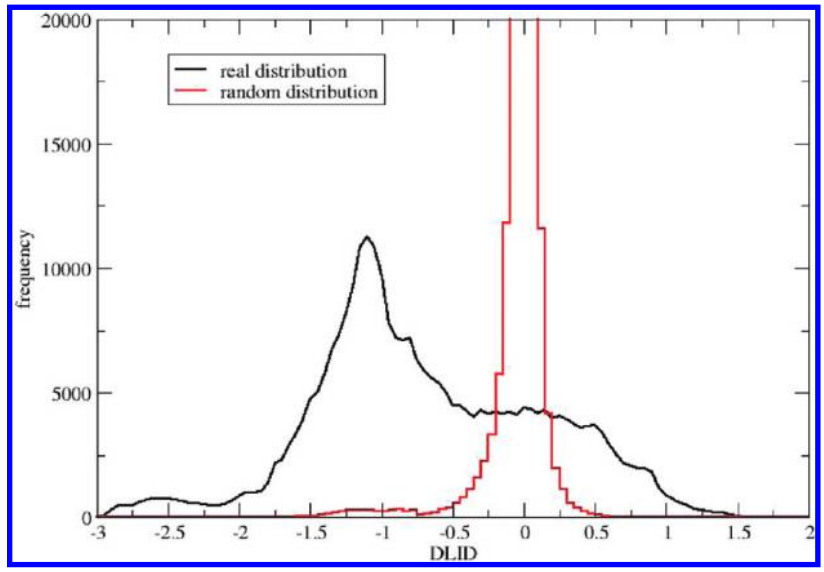

Figure 3. Histogram of DLID for all pockets given the observed distribution of drug-like ligand containing pockets in the three-dimensional pocket space shown in Figure 2. The line in red is for the case where the distribution of drug-like ligand containing pockets in pocket space is identical to the distribution of all pockets.

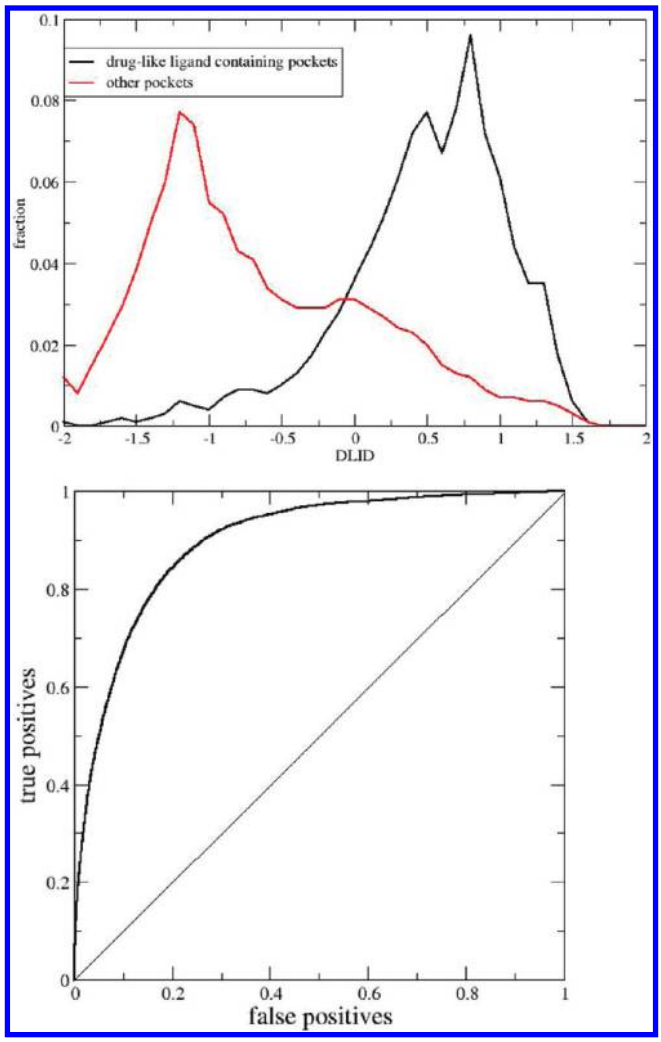

Figure 5. Top: Histogram for drug-like ligand containing pockets vs other pockets as a function of DLID. The histograms have been normalized for the same area under the curve. Bottom: ROC curve for finding drug-like ligand containing pockets as a function of decreasing DLID. The area under the ROC curve is 0.90.

pocket could be bindable if it had an especially large buriedness and/or hydrophobicity.

Discrimination of DLID. How well does the DLID distinguish between DLLC pockets and other pockets? Figure 5 shows the distribution of DLID for those two classes. Clearly DLID does not perfectly separate whether a pocket contains a drug-like molecule. Some discussion of the "mispredictions" is in order. It is expected that there should be many pockets that do not contain drug-like ligands but have high DLIDs. (These would be represented by the right side of the red curve in either plot in Figure 5, top.) One circumstance is that the pocket could potentially bind a drug-like ligand, but it just happens that no ligands were included in the crystallization. Another is that it does contain a ligand that is not drug-like. The converse, pockets that do contain a drug-like ligand but have low DLIDs (the left side of the black curve in Figure 5, top), is of more concern since these are potentially false negatives that would not be recognized as potentially drug-containing pockets based only on their pocket characteristics. Inspection of the more extreme cases shows that the reason the DLIDs are low is that the pockets are very small or shallow. The drug-like ligand associated with these pockets may be small (<15 non-hydrogen atoms) and fit inside the pocket. Other times the ligands are large (>20 non-hydrogens) and a large fraction of their atoms extend outside the pocket into "solvent", thus breaking the "50% of the atoms in the pocket" definition of drug-like. One could argue that pockets that do not enclose their ligands are less desirable and that having such false negatives is therefore not a large problem. Also, these targets could be rescued by looking at another chemical tractability metric—the availability of drug-like ligands—regardless of whether a crystal structure is available for the target−ligand complex. We examined the idea that drug-like ligands in the low-DLID pockets were disproportionately covalent binders, but this did not turn out to be the case.

The degree of separation in Figure 5 is typical of what we see in most types of virtual screening. We can apply the usual procedure for measuring virtual screening goodness, i.e., the retrospective screening experiment. If one sorts the pockets by decreasing DLID and counts the number of DLLC pockets found as a function of the fraction of total pockets examined, one sees an enrichment of 18−20 over random at 1% of the total pockets, where a random ordering of pockets would give an enrichment of 1.0 and a perfect ordering by DLID would give an enrichment of ∼50. Similarly one can generate a ROC curve (Figure 5, bottom) and find the area under the curve is ∼0.9, where random ordering would give an area of 0.5 and a perfect prediction would give an area of 1.0. These metrics are comparable to the enrichment of topological similarity searches,[44] which most cheminformatics workers consider good enough to be very useful, as long as the goal is to find many true positives (as opposed to avoiding false negatives).

Frequency of the Pocket with the Highest DLID Corresponding to the Pocket Containing the Most Drug-like Ligand. How often does the most bindable pocket as measured by DLID correspond to the pocket that actually contains the ligand with the lowest CI (i.e., the ligand with the most drug-like physical properties)? If we confine ourselves to the ∼5000 PDB entries that contain at least one drug-like ligand by our definition above and ignore residue types that occur >50 times in the PDB, the answer is 72%. Given that on average a PDB entry has ∼7 pockets, the "random" expectation for agreement would be ∼14%.

Inspection of the cases where the most bindable pocket by DLID is not the one containing the drug-like compound shows that the disagreement most often occurs in large proteins with many pockets (where the chance of there being another pocket with a more positive DLID is high) or where the drug-like ligand protrudes from a small or shallow pocket. The second is very much like the situation with pockets with low DLID that contain a drug-like ligand. Occasionally we have seen cases where the DLID of the pocket containing the drug-like ligand could not be calculated because there were too few neighbors in pocket space, and therefore another pocket appeared "best" by DLID.

Although DLID points to the pocket containing the drug-like ligand the majority of the time, we can certainly appreciate that the ligand-centric and pocket-centric views can tell very different stories about the bindability of a given PDB entry.

Best Pocket per PDB Entry. Since it only takes one good pocket to make a protein bindable, one necessary simplification for what follows is to take the pocket with the highest

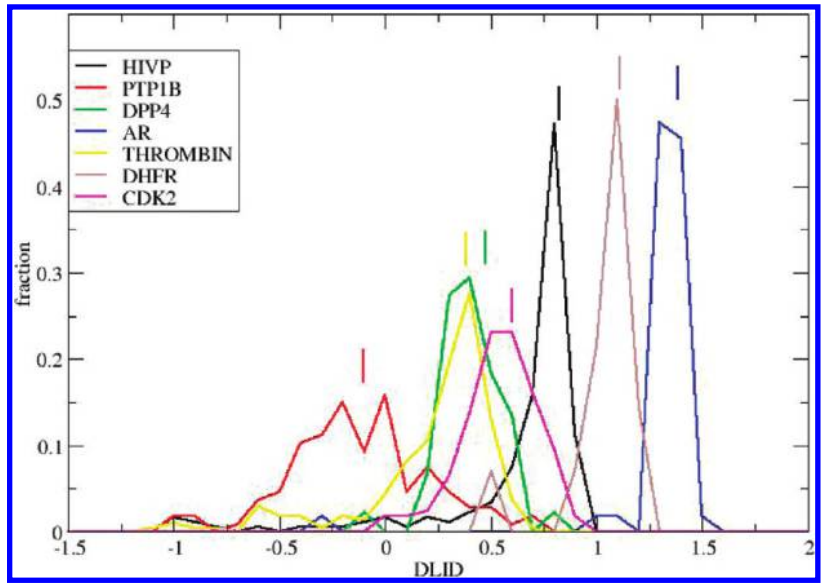

Figure 6. Distribution of DLIDs for some example targets based on clusters of similar sequences: HIV-1 protease (HIVP), human protein tyrosine phosphatase 1B (PTP1B), human dipeptidyl peptidase 4 (DPP4), human androgen receptor (AR), human thrombin (THROMBIN), pneumocystis dihydrofolate reductase (DHFR), and human CDK2 protein kinase (CDK2). The vertical bars indicate the median DLID for the individual clusters.

DLID as being representative of that crystal structure. We will refer to this as the "best pocket". This is often, but not always, the same pocket that contains an interesting ligand such as an inhibitor.

Variation of DLID within a Target and the Median DLID per Cluster. In an ideal world, the best pocket in different crystal structures of the same protein would have the same DLID. Our definition for the "same" protein is to use the 90% sequence identity clusters provided by the PDB, i.e., the same as our target definition. Inspection of individual clusters shows that in practice there are variations. The distribution of DLIDs for a few cases is shown in Figure 6. For HIV-1 protease (HIVP) the majority of best pockets for the crystal structures in the corresponding cluster have high DLIDs, but there are some crystal structures with anomalously low DLIDs. Some have the wrong biounits: a single chain only, such that the correct pocket cannot be formed. Others have a cocrystallized peptide longer than 7 residues; it was counted as part of the protein so that the normal pocket was not empty. Similarly, one can consider human PTP1B. Most of the crystal structures have low DLIDs, but there are a few cases for which the DLID could be considered in the bindable region. Inspection shows that these differences are due to real conformational shifts in protein loops. For the androgen receptor (AR), the anomalously low DLID for one crystal structure is due to the "correct" ligand-containing pocket having <30 neighbors in pocket space, thereby having no definable DLID, so that another pocket, one with a much poorer DLID, became best. Generally we find that artifacts that give anomalous DLIDs (missing residues, wrong biounits, etc.) are unavoidable in the large-scale automatic processing of the PDB. Even if we could eliminate artifacts by hand editing, there are still variations due to real changes (a loop moving, a side-chain filling a pocket, and a mutant residue, etc.). To define a single "official" DLID for a cluster, we take the median DLID over all the crystal structures associated with that cluster. This has the effect of "averaging out" the effect of variations, both real and artifactual. The mean DLID is too skewed by the anomalies, and therefore is a less desirable alternative compared to the median.

Level of Homologous Characteristics Needed for Targets To Have Similar DLIDs. In many cases a crystal structure is not available for the exact target of interest, but crystal structures may exist for a homologue. Can we draw conclusions about the DLID of the target based on the DLID of the homologous protein? For this we need to know how similar DLIDs for a target are as a function of the degree of homology. For a large sample of clusters where the number of PDB entries ≥ 3 we compiled the pairwise value:

$$|\mathrm{DLID}(i) - \mathrm{DLID}(j)|$$

where DLID(i) is the median DLID for all the PDB entries for cluster i. We also monitored the similarity of the sequences between clusters. There are potentially multiple sequences associated with each cluster, but since by definition they are ≥90% identical, it should suffice for our purposes to use the sequence of the first PDB entry in alphabetical order as representing that cluster. We monitored the $-\log(E)$ of all pairwise sequence comparisons of i and j, where E is the BLAST[45] probability that the sequences are unrelated. For comparison we also examined the variation of DLID within clusters, in which case DLID(i) would represent the best DLID for a particular crystal structure. This is a more comprehensive analysis of what was done in the above section. Figure 7 is a histogram of the distribution of the pairwise differences in median DLID. Most of the time the difference in median DLIDs between homologous clusters are small, clearly smaller on the average than for nonhomologous clusters. The degree of homology does seem to matter, more homologous clusters have more similar median DLIDs, but even pairs that would be borderline homologous ($4 < -\log(E) < 10$) are clearly different from nonhomologous clusters. The very homologous clusters ($-\log(E) > 50$) seem to have a similar profile compared to the variation of the DLIDs within a cluster, which is

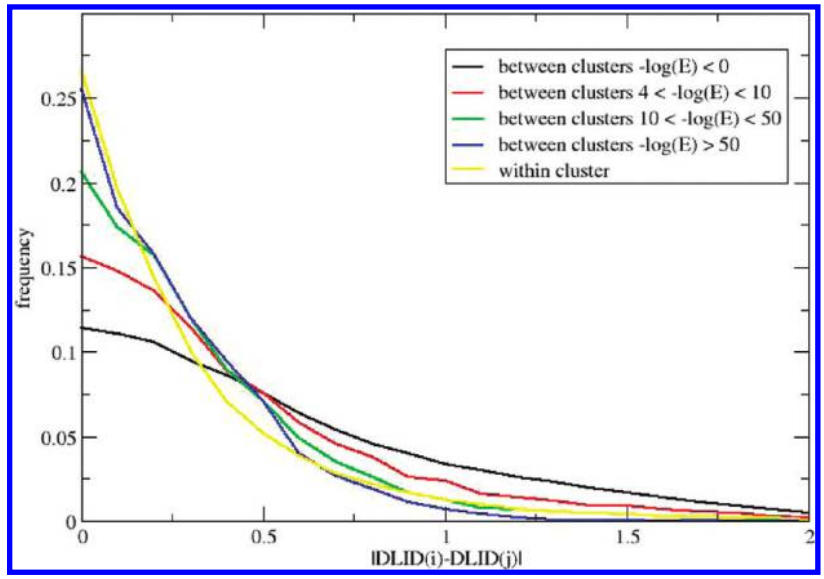

Figure 7. Histogram of the pairwise differences in DLID. In the case of "between cluster" pairs, we are comparing the median DLID of sequence clusters. In the case of "within cluster" we are comparing the best DLID of individual PDB entries in the same cluster. In both cases we are considering only those clusters with ≥3 PDB entries.

Table 1. Scoring Functions for Targets in Cheng et al.

Target	Source	Cluster number	Number of PDB entries	Median DLID[a]	MAP_{POD}[b]	mean Dscore[c]	Druggability score[d]
HIV INTEGRASE	HIV-1	7234	18	-0.22	18700	0.18	0.01
NEURAMINIDASE	Influenza A	2008	38	-0.12	2100	1.13	0.08
PTP1B	human	5512	108	-0.08	640	0.63	0.37
ANGIOTENSIN CONVERTING ENZYME	human	815	9	-0.04	130	1.00	0.39
CASPASE 1	human	6622	23	0.00	544	0.86	0.03
CATHEPSIN K	human	6142	31	0.09	150	0.77	0.30
FACTOR X	human	8754	69	0.11	61	1.07	0.12
DNA GYRASE B	E. coli	3563	3	0.27	1.5	1.03	0.43
THROMBIN	human	6526	218	0.37	5.3	1.10	0.51
EGFR PROTTEIN KINASE	human	4404	21	0.46	0.88	1.04	0.73
PENICILLIN BINDING PROTEIN	Streptococcus	437	8	0.56	230	1.04	0.25
INOSINE MONOPHOSPHATE DEHYDROGENASE	human	1387	4	0.60	190	0.88	0.04
CDK2 PROTEIN KINASE	human/bovine	6353	168	0.61	0.32	1.07	0.58
MAPK14 PROTEIN KINASE (P38)	human	3775	69	0.75	0.17	1.10	0.72
HIV PROTEASE	HIV-1	18759	201	0.81	0.66	1.06	0.64
FUNGAL CYP51	Mycobacterium	2216	9	0.85	0.35	1.24	0.62
COX2	mouse	765	8	0.86	0.022	1.25	0.72
ABL1 PROTEIN KINASE	human/mouse	1184	24	0.88	0.01	1.14	0.63
HMG CoA REDUCTASE	human	2032	22	0.88	61	1.06	0.12
ACETYLCHOLINESTERASE	Torpedo	851	68	0.94	0.53	1.16	0.61
ALDOSE REDUCTASE	human	5940	75	0.97	1.2	1.10	0.82
ENOYL REDUCTASE	human	8202	13	0.99	0.23	1.14	0.49
PHOSPHODIESTERASE 4D	human	3822	20	1.08	0.29	1.05	0.40
HIV RT (NNRTI)	HIV-1	963	103	1.18	0.0046	1.31	0.75
PHOSPHODIESTERASE 5A	human	4292	13	1.20	0.029	1.16	0.62

[a] Bindability based on median DLID: green, DLID ≥ 0.5; yellow, 0.2 ≤ DLID < 0.5; red, DLID < 0.2. [b] Druggability based on Cheng et al.:[15] green, druggable; yellow, difficult; red, undruggable. [c] From Halgren.[21] [d] From Schmidtke and Barril.[22]

expected. Collectively, these results indicate that one can transfer DLID between homologous targets.

DLID of Familiar Targets and Comparison with Other Scores. Having calibrated our bindability metrics using the presence of cocrystallized drug-like molecules, we would like to "validate" the metric against a set of targets with druggability assignments. We chose to look at those targets from Cheng et al.,[15] which most workers in this field take to be the baseline set. (Schmidtke and Barril[22] have started an effort to expand the number of targets so larger target sets will be available for future work.) The median DLIDs for clusters corresponding to those targets are listed in Table 1. We also include three other scores from the literature:

(1) MAP_{POD} from Cheng et al.[15] is a type of predicted binding constant (expressed in nM) based on simplified physical principles.

(2) Dscore from Halgren[21] is derived from SiteMap parameters. We list the mean Dscore over the crystal structures for each target examined by Halgren.

(3) A druggability score from Schmidtke and Barril[22] is based on fitting a function against physical parameters of

cavities for a large set of druggable and nondruggable cavities. We list the weighted mean over holo- and apo-forms of the targets given in the Supporting Information for that paper.

We note that the median DLID is based on all crystal structures in a cluster associated with a target and the Schmidtke and Barril druggability score is also based on a large number of crystal structures for each protein, whereas MAP_{POD} and Dscore are based on the same small number (1−6) of crystal structures for each target used by Cheng et al. (We found that the median DLID does not change significantly if we use only the few crystal structures used by Cheng et al.). We also note that MAP_{POD} defines binding sites in proteins on the basis of the distance to a cocrystallized ligand, whereas all the other scores define binding sites using algorithms that find concavities on protein surfaces.

Figure 8 shows the relationship between median DLID and the other scores. The Pearson correlation R of these other scores against median DLID is as follows: $-\log(MAP_{POD})$, 0.84; Dscore, 0.67; Schmidtke and Barril druggability score, 0.65. Clearly all scores in Table 1 are correlated, but why any particular score should track better or worse with median DLID is not clear since the methods differ widely in approach and implementation.

The color of the cells in Table 1 contains information about bindability/druggability assignments from median DLID and MAP_{POD}. Since median DLID tracks with MAP_{POD} and MAP_{POD} is a reasonable predictor for the Cheng et al. druggability assignments, it is not surprising that median DLID is mostly predictive of the Cheng et al. druggability assignments as well. Dscore and the Schmidtke and Barril druggability score have not been assigned a specific set of cutoffs from which to assign druggability, but one can see from Table 1 that the numbers generally track with the Cheng et al. assignments. If we set the Cheng assignments as "druggable" = 2, "difficult" = 1, and "undruggable" = 0, we can use the Pearson correlation of ranks (allowing for ties) to measure the ability of the scores in Table 1 to correctly order the Cheng assignments. The rank correlation values are as follows: median DLID = 0.76; $-\log(MAP_{POD})$ = 0.81; Dscore = 0.68; Schmidtke and Barril druggability score = 0.69. Thus all scores are roughly similar in this ability on this data set.

Level of Screen-out That Can Be Obtained Using the DLID Metric. For pocket characteristics to be useful in practice, we need to be able to eliminate some large fraction of the PDB entries as attractive targets. If, for example, 90% of PDB entries had at least one bindable pocket, pocket characteristics would be practically worthless because very little could be eliminated. So what fraction of PDB entries is bindable? Using the DLID cutoff of 0.5 defined above, only 8.4% of all ~292 000 pockets have a DLID ≥ 0.5, i.e., are bindable. Going to the next level, the best DLID per PDB entry, we see that 36% of the ~42 000 PDB entries are bindable. At the highest level, 38% of the median DLIDs for clusters (where there are at least 3 PDB entries) ≥ 0.5. Thus, using a DLID cutoff of 0.5, no more than a minority of pockets, PDB entries, or targets are bindable, and a significant fraction of targets can be eliminated using pocket characteristics. A histogram of DLID for the above levels is shown in Figure 9.

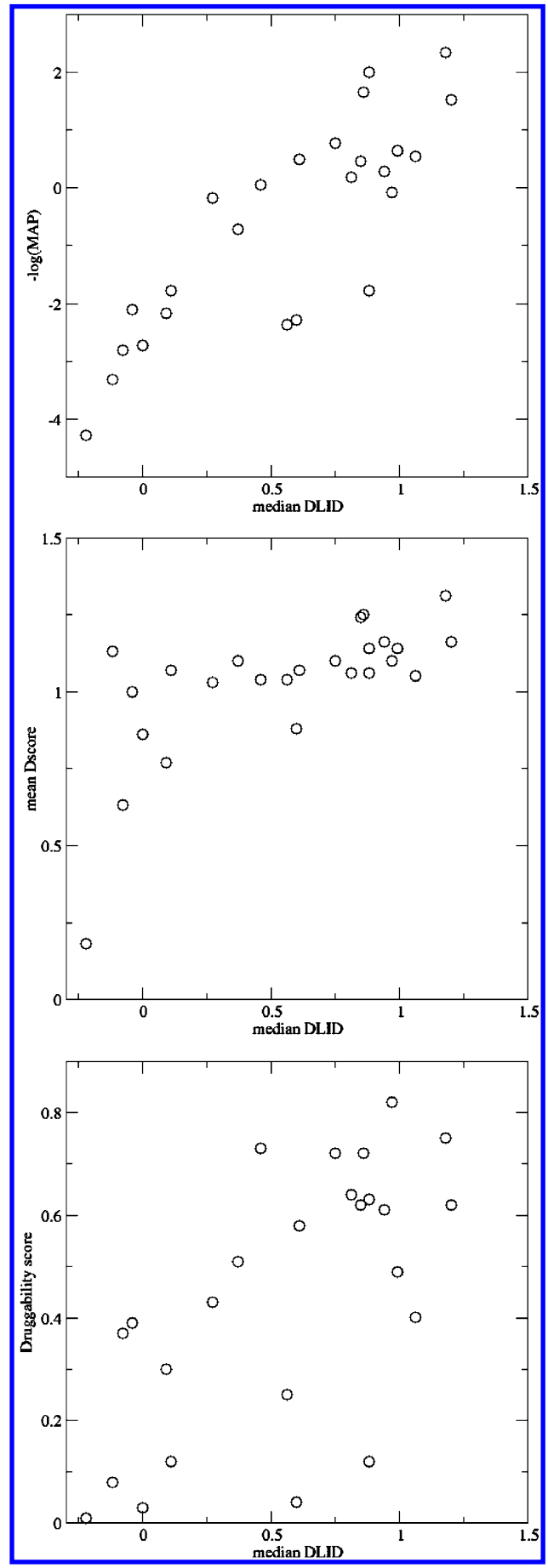

Figure 8. Comparing median DLID (this work) to other scoring functions in Table 1 for a set of targets from Cheng et al.

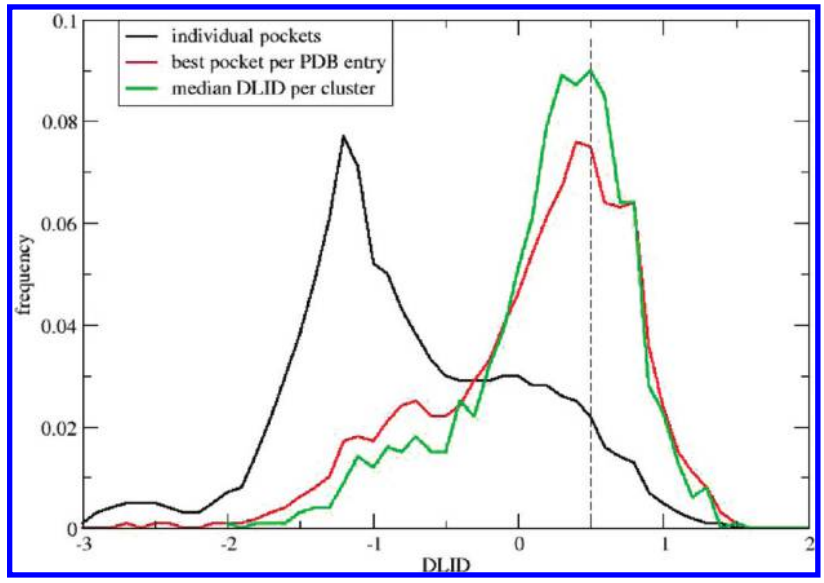

Figure 9. Histogram of the occurrence of DLID for different levels of data agglomerations: individual pockets, best pocket per PDB, and median DLID per cluster where there are at least 3 PDB entries/cluster. The dashed line indicates our choice for a DLID cutoff above which we consider a pocket bindable.

DISCUSSION

We have invented a simple metric DLID for measuring the bindability of proteins that depends on the volume, buriedness, and hydrophobicity of pockets. All of these characteristics are easy to understand and calculate, and it is easy to visualize the pocket space consisting of these parameters. While others have settled on similar properties in their studies, ours is the most comprehensive in terms of varieties of crystal structures examined. Generally speaking the qualitative result is not a surprise in retrospect. We know the ligand-binding sites of proteins are large (volume) surface grooves or invaginations (buriedness) that are less polar (hydrophobicity) than the overall surface. However, it is possible to quantify the individual contributions of these parameters as to the relative ability of a pocket to bind a drug-like ligand. It should be noted in this context that Schmidtke and Barril[22] have indicated that the arrangement of polar groups in a binding site, not just the overall hydrophobicity, could be important for druggability, something we have not addressed here.

The DLID approach is "agnostic" in the sense that it depends only on selecting drug-like molecules out of the universe of ligands cocrystallized with proteins. It is not calibrated against a previous determination of whether a particular target is druggable,[22] which can be arbitrary and/or subjective, or an experimental determination of whether fragments bind to the target,[12] which is resource-intensive. The advantage is that we are not confined to a small set of proteins (a few dozen) but can use the majority of the proteins in the PDB as a source of data. The drawback is that we cannot easily consider the potency of binding because even large databases such as MOAD[43] do not contain binding constants for more than a small fraction of the cocrystallized ligands.

One should note that although protein−protein interfaces are of interest as drug targets, our methodology has difficulty addressing them. In the complexed form of two interacting proteins, the ligand is a large protein which is not removed, so the interface is inaccessible. In the crystal structures of the individual proteins, the protein−protein interface is often so flat that icmPocketFinder does not perceive a concavity. Only in the crystal structures where the interface of one of the proteins is complexed with a small molecule that induces a deeper pocket is the pocket perceived.

There are several important caveats about the PDB that are relevant to the bindability issue. Two have to do with the proteins that appear in the PDB:

(1) The PDB is heavily biased toward therapeutically relevant targets (otherwise why bother to elucidate their structures?), so poor targets are underrepresented. On the other hand, some good targets are also underrepresented because they are hard to crystallize. In particular, class A GPCRs, one of the most important classes of drug targets, is represented by only a dozen or so structures, and the majority of them are opsins rather than receptors.

(2) There is never a guarantee that what appears in the crystal structure is the species to which a drug-like compound would actually bind. For example, consider the case of HIV integrase. While there is a very successful drug raltegravir that binds to that target, all pockets are highly undruggable/unbindable by all scoring methods in Table 1. The proper drug-binding species for the HIV integrase strand transfer inhibitors is probably protein plus viral DNA,[46] whereas all the crystal structures contain only protein. Recent crystal structures of the related prototype foamy virus integrase cocrystallized with raltegravir and elvitegravir clearly demonstrate this binding mode.[47]

Another set of caveats has to do with the lack of standardization/notation in the PDB that are relevant when one has to deal with a very large number of structures (10s of thousands), rather than a small (a few dozen) hand-curated sets of binding sites that already contain drug-like ligands:

(3) Although the PDB has recently undergone a global "remediation" that greatly improved the consistency among PDB entries, there is still significant inconsistency from author to author on a number of aspects (how biounits are represented, whether ligands are represented as one or many residues, whether active sites are noted by the author, etc.).

It is hard to formulate an automated workflow such that all entries are treated correctly, and the scale is too large for a manual fix-up.

(4) The PDB has no explicit method of indicating which heterogroups are "ligands" vs "cofactors". This makes it almost impossible to include cofactors other than heme as part of the protein when calculating pockets without an impractical level of hand annotation.

A third set of caveats has to do with variation among different crystal structures of the same target protein such that one cannot take a single crystal structure of that target as representative of all of them as far as binding site characteristics go. This has not generally been addressed until recently, for example, by Schmidtke and Barril.[22]

(5) Sometimes side chains, and even entire sets of residues, are missing in some PDB entries of the same protein but not others.

(6) Some crystal structures are of mutants. Sometimes the mutation is made to form better crystals. Sometimes the mutation is made to judge the effect on the binding of ligands. Mutated side chains can change the size and nature of binding sites.

(7) Proteins are surprisingly flexible. This means that different instances of the identical protein can have different binding sites: a large binding site in one structure can be split into two smaller sites if a side chain blocks the center, a ligand can induce a conformational change in the protein that makes a binding site much larger, etc.

Relating genes to pocket information is a multilevel problem. The goal is to start with a list of gene sequences and look for homologous proteins in the PDB that have "good pockets". There can be multiple homologues of the gene in the PDB with various degrees of homology, and a number of PDB entries may represent the same protein. A PDB entry can be interesting or not by a number of criteria (resolution, for example). Each PDB entry has a number of pockets, and each pocket has its own DLID. Hence we need to make at least two simplifying assumptions. First we are interested only in the highest DLID for a given crystal structure. Second, we are using the median highest DLID over all crystal structures to be the DLID of that protein. The problem therefore reduces to finding targets that are simultaneously most homologous to the gene and which have the highest median DLID. As with any data-mining exercise on a large, noisy data set that one cannot easily hand-curate, one is required to inspect the "final positives" one has selected to see if they are truly positive. Fortunately, for the purposes of identifying bindable targets, one need not find every positive; having many positives is probably sufficient.

Despite the complications discussed above, we can answer the following questions:

(1) Is it possible to identify pockets likely to contain drug-like ligands from their physical properties given the data that exists in the PDB? Yes, most of the time.

(2) Is it possible to use pocket information to help prioritize target genes? Yes, to a useful extent.

We must add a caveat reminding the reader that bindability is not totally predictive of a target being successfully addressed in a drug discovery program. Low DLID implies only that the active site may be very small, or shallow, or very polar. Therefore it may be difficult to develop a high affinity small noncovalent ligand in conventional ways. However, as the successful drug targets, Cathepsin K and HCV protease, have demonstrated, one may use covalent ligands such as boceprevir or telaprevir or very large ligands such as vaniprevir, respectively, to achieve potency even though the DLID indicates a very shallow unbindable pocket. Similarly, DPP4 and BACE have DLID values at the level of 0.45−0.46, on the borderline between bindable and difficult because of their polar binding sites. DPP4 has a successful inhibitor, sitagliptin doing well on the market, while BACE has been very challenging as a drug target. One of the differences between them is the locations of the enzymes, which require different physical properties of the inhibitors; BACE inhibitors must be brain penetrant for achieving efficacy which is not required for sitagliptin type of DPP-4 inhibitors which need bind only to cell surfaces.

ACKNOWLEDGMENT

We are very grateful to all our modeling colleagues for providing information and insights on the targets they are supporting, in particular, John Sanders in West Point and Andreas Verras and Qioalin Deng in Rahway for detailed discussions on their targets. We also thank Chris Culberson in West Point for his support on the MIX command frompdb which made it possible for us to process the PDB effectively.

Supporting Information Available: List of pockets, their characteristics, and DLID, list of clusters, the PDB entries in their clusters, and DLID; list of clusters, the PDB entries in the clusters, and the median DLID for the clusters; and structures of "drug-like" ligands. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Russ, A.; Lampel, S. The druggable genome: An update. *Drug Discovery Today* **2005**, *10*, 1607–1610.
(2) Zheng, C. J.; Han, L. Y.; Yap, C. W.; Xie, B.; Chen, Y. Z. Trends in exploration of therapeutic targets. *Drug News Perspect.* **2005**, *18*, 109–127.
(3) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there. *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.
(4) Imming, P.; Sinning, C.; Meyer, A. Drugs, their targets and the nature of number of drug targets. *Nat. Rev. Drug Discovery* **2007**, *5*, 821–834.
(5) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.
(6) Cheng, A. C. Predicting selectivity and druggability in drug discovery. *Annu. Rep. Comput. Chem.* **2008**, *4*, 23–37.
(7) Landry, Y.; Gies, J.-P. Drugs and their molecular targets. *Fundam. Clin. Pharmacol.* **2008**, *22*, 1–18.
(8) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* **2009**, *14*, 155–161.
(9) Harland, L.; Gaulton, A. Drug target central. *Expert. Opin. Drug Discovery* **2009**, *4*, 857–872.
(10) Han, L. Y.; Zheng, C. J.; Xie, B.; Jia, J.; Ma, X. H.; Zhu, F.; Lin, H. H.; Chen, X.; Chen, Y. Z. Support vector machines approach for predicting druggable proteins: Recent progress in its exploration and investigation of its usefulness. *Drug Discovery Today* **2007**, *12*, 304–313.
(11) Sakharkar, M. K.; Sakharkar, K. R.; Pervaiz, S. Druggability of human disease genes. *Int. J. Biochem. Cell Biol.* **2007**, *39*, 1156–1164.
(12) 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.
(13) Brown, S. P. Hajduk PJ Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* **2006**, *1*, 70–72.
(14) Coleman, R. G.; Salzberg, A. C.; Cheng, A. C. Structure-based identification of small molecule binding sites using a free energy model. *J. Chem. Inf. Model.* **2006**, *46*, 2631–2637.

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

(16) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.* **2007**, *47*, 2287–2292.

(17) Hambly, K.; Danzer, J.; Muskal, S.; Debe, D. A. Interrogating the druggable genome with structural informatics. *Mol. Diversity* **2006**, *10*, 273–281.

(18) Weisel, M.; Proschak, E.; Kriegl, J. M.; Schneider, G. Form follows function: Shape analysis of protein cavities for receptor-based drug design. *Proteomics* **2009**, *9*, 451–459.

(19) Gupta, A.; Gupta, A. K.; Seshadri, K. Structural models in the assessment of protein druggability based on HTS data. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 583–592.

(20) Seco, J.; Luque, F. J.; Barril, X. Binding site detection and druggability index from first principles. *J. Med. Chem.* **2009**, *52*, 2363–2371.

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

(22) Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858–5867.

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

(24) Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K. Quantifying the relationship among drug classes. *J. Chem. Inf. Model.* **2008**, *48*, 755–765.

(25) Kuhn, M.; Campillos, M.; González, P.; Jensen, L. J.; Bork, P. Large scale prediction of drug-target relationships. *FEBS Lett.* **2008**, *582*, 1283–1290.

(26) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Cambell, R. M. Kinomics: characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839–846.

(27) 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.

(28) An, J.; Totrov, M.; Abagayan; et al. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inf.* **2004**, *15*, 31–41.

(29) Zhong, S.; MacKerrel, A. D., Jr. Binding response: A descriptor for selecting ligand binding site on protein surface. *J. Chem. Inf. Model.* **2007**, *47*, 2303–2315.

(30) Halgren, T. A. New method for fast and accurate binding-side identification and analysis. *Chem. Biol. Drug Des.* **2007**, *69*, 146–148.

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

(32) Nicola, G.; Smith, C. A.; Abagyan, R. New method for the assessment of all drug-like pockets across a structural genome. *J. Comput. Biol.* **2008**, *15*, 231–240.

(33) Ghersi, D.; Sanchez, R. EasyMIFS and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **2009**, *25*, 3185–3186.

(34) Weskamp, N.; Hullermeier, E.; Klebe, G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins* **2009**, *76*, 317–330.

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

(36) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2010**, *23*, 209–219.

(37) Campagna-Slater, V.; Arrowsmith, A. G.; Zhao, Y.; Schapira, M. Pharmacophore screening of the Protein Data Bank for specific binding site chemistry. *J. Chem. Inf. Model.* **2010**, *50*, 358–367.

(38) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

(39) ICM-pro, Version 3.0, Molsoft, LaJolla, CA, http://www.molsoft.com.

(40) Bush, B. L.; Sheridan, R. P. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.

(41) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer automated log *P* calculations based on an extended group-contribution approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.

(42) The Physicians Desk Reference, http://www.pdr.net (accessed Sep. 1, 2010.

(43) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* **2008**, *36*, D674–D678.

(44) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

(45) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids. Res.* **1997**, *25*, 3389–3402.

(46) Espeseth, A. S.; Felock, P.; Wolfe, A.; Witmer, M.; Grobler, J.; Anthony, N.; Egbertson, M.; Melamed, J. Y.; Young, S.; Hamil, T.; Cole, J. L.; Hazuda, D. J. HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique strand transfer conformation for integrase. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11244–11249.

(47) Hare, S.; Gupta, S. S.; Valkov, E.; Engleman, A.; Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **2010**, *464*, 232–237.

CI100312T