Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins

Paul S. Charifson,* Joseph J. Corkery, Mark A. Murcko, and W. Patrick Walters

Vertex Pharmaceuticals, 130 Waverly Street, Cambridge, Massachusetts 02139-4242

Received July 8, 1999

We present the results of an extensive computational study in which we show that combining scoring functions in an intersection-based consensus approach results in an enhancement in the ability to discriminate between active and inactive enzyme inhibitors. This is illustrated in the context of docking collections of three-dimensional structures into three different enzymes of pharmaceutical interest: p38 MAP kinase, inosine monophosphate dehydrogenase, and HIV protease. An analysis of two different docking methods and thirteen scoring functions provides insights into which functions perform well, both singly and in combination. Our data shows that consensus scoring further provides a dramatic reduction in the number of false positives identified by individual scoring functions, thus leading to a significant enhancement in hit-rates.

Introduction

Over the past 10 years, the use of various docking methods has become commonplace to identify leads from compound collections when structural information of the target has been determined.[1,2] Many of these same docking algorithms have also been applied toward structure-based library design in both lead identification and lead optimization contexts. In addition to the necessity of possessing the desired target's three-dimensional coordinates as well as a docking method, both structure-based database mining and library design also require a virtual collection of desired test compounds and one or more methods to rank these compounds. The scoring functions used to rank compounds should be able to distinguish active from inactive compounds independent of the docking method used and should be comprised of physically intuitive and interpretable terms. However, it does not necessarily follow that these same scoring functions should be able to predict binding affinity in statistically rigorous terms since the functional forms used to describe the chemistry and physics of ligand binding are notoriously incomplete.[3] The majority of published scoring functions have been developed in association with docking methods. These docking methods have typically been validated on the basis of their ability to reproduce the geometries of high-affinity protein−ligand complexes. While this is a necessary criterion, it does not address one of the primary uses of a docking program, the identification of novel micromolar lead compounds. In many cases, the scoring functions that have evolved with these docking methods can consistently predict the binding mode of ligands binding with nanomolar affinity; however, they perform poorly for predicting lower affinity binders. This is especially true for attempts to identify low affinity ligands for systems outside the training sets of some empirical scoring functions.

In this work, we present an evaluation of two different docking methods and thirteen different scoring functions as applied to three current targets of high pharmaceutical interest. The motivation for this study came from the observation that we were consistently obtaining confirmed hit-rates ($IC_{50} < 50\ \mu M$) between 2 and 7% for a variety of enzyme targets, while screening relatively small numbers of compounds (30−400). These hits, representing multiple compound classes, came from commercially available compound sources and were obtained with DOCK 4.0.1 employing several scoring functions in an approach we will describe as consensus scoring. We wished to determine whether the scoring functions we had been using were optimal for the docking/ lead identification process and whether scoring function performance was generalizable and independent of the docking method. Our desire to optimize performance in this area stems from our overall screening philosophy in which compounds are selected by a variety of different methods for directed screening (e.g., both 2D and 3D similarity, diversity, and docking).

Database Integrity and Consensus Scoring. We believe that the consistently better than random confirmed hit-rates we had observed can be attributed to two factors. First, we have been careful to ensure that any database we were docking had been filtered by REOS [4] to remove compounds containing functional groups known to be reactive or resemblant of toxic entities. We further ensured that all compounds were within ranges of properties calculated for known drugs in a manner similar to Lipinski et al.[5] These simple filters serve to reduce the incidence of false positives, improve the downstream properties of the compounds (pharmacokinetics, metabolism), and provide "drug-like" leads from the outset. The second reason we believe we have obtained our observed hit-rates is due to the way we apply scoring functions during and after the docking process. We routinely combine several scoring functions in an intersection approach toward compound selection.

* Corresponding author. Phone: 617-577-6442. Fax: 617-577-6680. E-mail: paulc@vpharm.com.

10.1021/jm990352k CCC: $18.00 © 1999 American Chemical Society
Published on Web 11/25/1999

Approaches of this type have been have been applied in the areas of QSAR[6] and molecular similarity.[7,8] This consensus scoring approach simply involves scoring compounds with several methods (typically 2−3) and taking the intersection of the top *N*% of each of these sorted lists. In practice, this is accomplished by use of a primary scoring function followed by rescoring of the best configuration (i.e. orientation/conformation) identified during docking with the other functions. Another goal of this study was to, therefore, determine, if there were optimal combinations of functions that could be identified for use by a consensus approach. We also wished to determine to what extent the primary function influenced the overall results.

Docking Methods and Scoring Functions. There are presently several available docking programs such as DOCK,[9−11] Autodock,[12−15] FlexX,[16−18] GOLD,[19,20] and Flexdock[21] that solve the docking problem and the conformational flexibility issues in slightly different ways. For the purposes of this study, we have evaluated DOCK and a Genetic Algorithm-based docking method of our own (see Methods section). Although there are capabilities in DOCK 4.01 to perform flexible docking, we have chosen to evaluate DOCK in the traditional site point/internal distance matching paradigm, since this is the most commonly used docking approach. Multiple conformers (see Methods section) have been employed in this study; we[4,22] and others[23] have found this to be an efficient way to address the conformational flexibility issues. In a similar fashion, our GA-based docking program, GAMBLER, was used simply to orient multiple conformers (i.e., the GA was not applied to searching torsions).

Thirteen scoring functions are considered in this study and are loosely grouped into three categories (Figure 1): empirical functions, molecular mechanics-based functions, and functions not cleanly fitting into one of the previous two categories. It is beyond the scope of this paper to review all of these functions, thus we will comment only on specific performance as it applies to this study. Since we did not have access to all of these programs/algorithms, we implemented the following functions: Böhm,[24] ChemScore,[25,26] Piecewise Linear Potential (PLP),[27] FLOG,[28] Volume Overlap.[29] In every case, we have attempted to reproduce test cases presented by the original authors. In some cases, more than one version of a given algorithm has been published and we have given the appropriate reference for the version we used. Additionally, we have employed simplex[30] minimizations for the following functions: ChemScore, PLP, DOCK (energy, chemical, contact) scores, and FLOG. The MMFF[31−33] nonbonded interaction energies and strain energies involved conjugate gradient minimizations. Details for some of the scoring functions are provided in the Methods section.

Study Design. Three targets for which the high-resolution crystallographic structures are known were used in this study. These include p38 MAP kinase,[34] inosine monophosphate dehydrogenase (IMPDH),[35,36] and HIV protease.[37−39] Although all three of these sites are reasonably buried, each system represents the practical complexities encountered in docking calculations. P38 has been shown to exhibit subtle, yet significant rearrangements in the glycine rich loop of the ATP binding site when binding chemically different inhibitor classes. IMPDH also exhibits some conformational mobility in a region of one of the flaps comprising the active site and also has a cofactor requirement. HIV protease has a conserved flap water that can be considered as part of the protein or removed to allow occupancy from a potential inhibitor atom. Additionally for HIV protease, the natural ligand is a peptide and the identification of nonpeptide inhibitors for proteases has been a challenging problem.

Empirical

Böhm (bohm) [24]

ChemScore (csco) [25,26]

SCORE (score) [58]

Piecewise Linear Potential (plp) [27]

Molecular-Mechanics

Merck Molecular Force Field non-bond energy (nbt) [31-33]

DOCK energy score (nrg) [9,11]

DOCK chemical score (chm)

Flexible Ligands On a Grid (flog) [28]

Strain Energy (strain)

Other

Poisson-Boltzman (pb) [49]

Buried Lipophilic Surface Area (bsa) [59]

DOCK contact score (cnt) [60]

Volume Overlap (vov) [29]

* labels in parentheses are used in Figures 2 and 3

Figure 1. Classification of scoring functions evaluated in this study.

Table 1. Number of Unique Molecules and Conformers for Each of the Study Sets

target	rot. bond cutoff	#unique molecules	number conformers
P38	7	502	21691
IMPDH	7	400	66761
HIV protease	12	787	112131

The test compounds for each target came from Vertex research programs. In each case, high quality experimental inhibition constants have been measured, the K_i or K_i apparent ranged from low nanomolar to micromolar. For each target, the number of test compounds used in the study was further reduced by applying a rotatable bond cutoff. This also serves to improve the chances that we are generating conformers that might be close to the necessary binding conformation. Table 1 shows the rotatable bond cutoffs employed as well as the number of unique test compounds and the number of multiple conformers generated for each compound. Additionally, 10 000 randomly chosen compounds, which had previously been subjected to REOS filters, were selected from commercial sources and subjected to the

Scheme 1

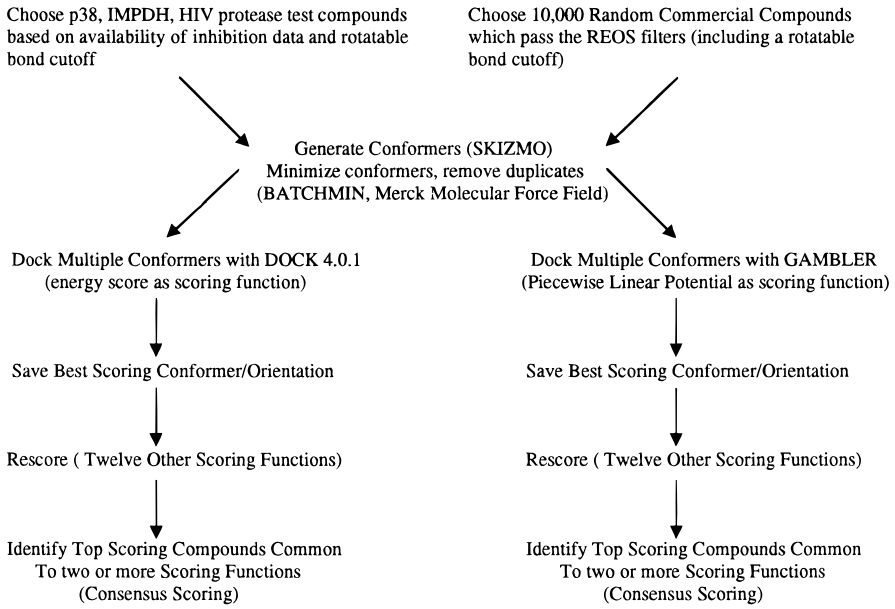


Table 2. Number of Compounds in Each of Three Activity Ranges for Each Target

target	<100 nM	100 nM − 1.0 μM	1.0−30.0 μM
P38	126	119	256
IMPDH	172	137	91
HIV protease	275	146	360

same conformer generation techniques (see Methods section) resulting in 189 378 conformers.

We further partitioned each of the study sets for each target into three activity ranges based on inhibition constants: less than 100 nM, between 100 nM and 1.0 μM, and greater than 1.0 μM, but less than 30.0 μM. Table 2 shows the number of compounds in each of these activity ranges for each study system. If we assume that the randomly chosen commercial compounds are "inactive", then the basic paradigm for this study is described by docking multiple conformers of the corresponding test compounds along with multiple conformers of the 10 000 random compounds into each target. We then ask to what extent can we correctly identify the active compounds in a given activity range (i.e., from the 10 000 random inactive compounds + active test compounds) using various combinations of one, two, or three scoring functions? As mentioned above, we use an initial scoring function for each docking method. We then save the best docked orientation of only the best scoring conformer of each molecule and then rescore that configuration with the other twelve scoring functions. Scheme 1 shows the basic flow of this study.

Methods

Conformer Generation. Conformers were generated using SKIZMO[40] in a gas-phase calculation in which duplicates (including symmetry-related duplicates) were removed (within 0.5 Å RMSD) after conformer generation. Conformers were then subjected to energy minimization in BATCHMIN[41] using the MMFF force field with the GB/SA continuum solvation model[42] to a derivative convergence of 0.01 kJ/Å/mol, if obtainable within 1000 steps of conjugate gradient minimization. All conformers within 6.0 kcal/mol of the lowest energy conformer were retained and duplicates were again removed from the minimized set of conformers.

DOCK 4.0.1. Docking spheres were generated from crystallographic nonhydrogen atom positions of inhibitor atoms derived from the appropriate cocomplexes. Four sphere types were employed (donor, acceptor, polar, hydrophobic) and critical cluster spheres were assigned to at least one key hydrogen bonding position: p38 (45 spheres total from two crystallographic complexes; 10 typed spheres, 1 critical cluster sphere), IMPDH (45 spheres total from two crystallographic complexes; 6 typed spheres, 2 critical cluster spheres), HIV protease (35 spheres total from one crystallographic complex; 6 typed spheres, 2 critical cluster spheres). Energy scoring employing rigid-body simplex minimization was utilized as the primary scoring function. Automated matching was used allowing up to 500 orientations per conformer to be examined. Gasteiger-Marsili[43] charges were loaded for all conformers by BABEL[44] for the calculation of the electrostatic component of the energy scores. Docking grids were generated by the grid utility program within the DOCK 4.0.1 distribution using default values. The box dimensions used to calculate the grids were: p38 (22.8 × 24.3 × 23.2), IMPDH (27.6 × 20.1 × 18.8), HIV protease (23.0 × 21.7 × 29.5).

GAMBLER. Docking was also carried out using an internally developed genetic algorithm based docking program known as GAMBLER (Genetic Algorithm Multiprocessor Box-Oriented Ligand Enzyme Relocator). The GAMBLER program follows a model utilized in a number of previously developed programs.[27,45−47] GAMBLER begins with a random population of chromosomes that encode translation and rotation matrices. Each chromosome in this random population is evaluated by applying the translation and rotation matrices and positioning the ligand in the active site. Each chromosome is then assigned a fitness score according to the value calculated by the scoring function. High scoring chromosomes are combined to produce a new population in an operation known as crossover. To maintain diversity in the population and allow ligands to explore new regions of space, random mutations are also performed on a percentage of the population. The docking region is defined by a rectilinear box. Sampling efficiency is increased by initially subdividing the active site into eight sub-boxes and searching each sub-box.[45] The highest scoring chromosomes from the sub-boxes are then used to form the initial population. The initial coarse search performed in each sub-box utilized a population size of 250 for 12 generations. From the total set of 3000 solutions generated, the best 250 were selected and searched further for 100 generations. The PLP function was used as the primary scoring function to determine which configuration would be saved for rescoring.

Prior to docking either the appropriate test compounds or

randomly selected commercial compounds, control docking calculations were performed to ensure that the original crystallographic complexes could be reproduced with the docking methods/primary scoring functions employed. For DOCK (using site points selected from the same crystallographic complex ligand nonhydrogen atoms), all three complexes could be reproduced to within 0.5 Å rmsd of the experimental complex while for GAMBLER (with no site points), all three experimental complexes could be reproduced to within 1.0 Å rmsd.

Scoring Function Details. The implementation of the functions came from the published work (Figure 1). Only details specifically pertaining to this study are discussed: due to instabilities in the simplex minimizations on FLOG grids, we performed each simplex 10 times and report the median values of the 10 calculations. Nonbonded interaction energies calculated with MMFF from BATCHMIN were implemented through the use of Perl[48] scripts. Conjugate gradient minimization proceeded for 250 steps keeping the protein atoms completely immobile. Strain energy calculations also utilized MMFF/BATCHMIN from Perl scripts and were carried out by an initial restrained minimization to the docked geometry (half-width of flat bottom restraint = 0.5 Å, force constant = 500 kcal/mol/Å^2) to convergence (0.01 kJ/Å/mol) followed by removal of the constraints and full minimization until convergence (0.01 kJ/Å/mol) into the closest local minimum. PLP scores reported in all figures are the result of rescoring against the PLP grids with simplex minimization (the initial GAMBLER dockings used PLP without minimization as a primary scoring function). ChemScore was implemented as a continuous function, not yet implemented on a grid. For the Poisson−Boltzmann calculations we employed MMFF charges and radii from GRASP.[49,50] Volume overlap calculations were performed relative to the original crystallographic ligand and were implemented in a grid-based occupancy approach.

Results and Discussion

Although the focus of this study is to determine the utility of combining scoring functions in a consensus approach, it is instructive to evaluate the individual performance of the functions considered. In this context, performance is defined as the percent active compounds returned by the docking experiment as we evaluate successively more and more of all top scoring compounds (i.e., actives and inactives). Figure 2, A−D, illustrate the performance of each scoring function for each of the three activity ranges against each of the three study systems. In each case, the score for a given molecule is assigned based on that of its highest scoring conformation. Figure 2, A, C, and D, represent the results with DOCK for p38, IMPDH, and HIV protease, respectively, while Figure 2B shows the GAMBLER results for p38. For each of these plots, the *x*-axis denotes the top thousand scoring molecules (approximately 10%) considered for a given function including both test compounds and randomly chosen (inactive) commercial compounds, while the *y*-axis denotes the percentage of "active" test compounds correctly identified in that activity range. In each figure, the top panel represents the nanomolar compounds, the middle panel represents the 100 nM compounds and the bottom panel represents the micromolar compounds.

Figure 2A shows that some of the functions are quite effective at identifying the nanomolar p38 compounds relative to the presumed inactive commercial compounds. This is reassuring, but not impressive since, in practice, it is extremely rare to find nanomolar compounds in screening databases of commercially available compounds; most high throughput screening endeavors identify mainly micromolar compounds.[51] The better performing functions in this example include ChemScore, DOCK energy score, FLOG, PLP, and the SCORE function. Typically, we find 50% of the nanomolar hits in the top 5−10% of the total compounds considered (i.e., actives + random). It is interesting to note that the plot plateaus at less than one hundred percent illustrating the general limitations in the docking and scoring functions examined in this study. The middle panel of Figure 2A, representing the 100 nM compounds, shows a similar relative ordering of the same good performing functions, but the absolute number of actives correctly identified is reduced, overall. The bottom panel of Figure 2A shows a marked compression of the function performance as would be expected for micromolar compounds, but it is once again interesting to note that the same functions are outperforming the other functions, although in this case volume overlap performs reasonably well. The GAMBLER docking/rescoring for the p38 system (Figure 2B) is quite similar in function performance to those produced by DOCK and the same good performers (i.e., ChemScore, DOCK energy, PLP, and FLOG) appear to be maintained. In this case, the DOCK chemical score also appears to perform reasonably well along with the SCORE function. In general, since the results for DOCK and GAMBLER are quite similar, the GAMBLER results for IMPDH and HIV protease are not shown. It is encouraging that in the GAMBLER dockings a different primary scoring function was used, but upon rescoring the same good performing functions are consistent with those observed with DOCK.

For the IMPDH system (Figure 2C) we see an enhanced ability to correctly identify active compounds relative to the more difficult p38 case in which larger changes in protein conformation are observed upon binding different inhibitor classes. The best performing functions are, again, the ChemScore function, DOCK energy score, and PLP. The DOCK contact score also performs reasonably well in this system. As previously noted for p38, similar trends are observed for the nanomolar and 100 nM compounds. For the micromolar compounds, we once again see a reduced performance, but with the ChemScore function, DOCK energy score, and PLP doing well. In this case, the Böhm function and volume overlap also appear to perform reasonably well. Visual inspection of docked structures suggests that the exceptional performance of the top scoring functions in this case is overestimated (i.e., that some incorrectly docked IMPDH inhibitors are scoring better than they should).

In the case of HIV protease (Figure 2D), the functions best able to pull out the correct active compounds from inactive compounds are: ChemScore, DOCK energy score, PLP, and DOCK contact score. This trend is upheld across the three activity ranges and shows good performance in the micromolar range as well. Visual inspection of docked structures in this instance supports that the majority of docked compounds are correctly docked and that the functions are performing well. However, there are several factors influencing the good results we see in this case. First, in the case of the exceptional ChemScore performance across the three activity ranges, we must not forget that several HIV protease inhibitor complexes were included in the

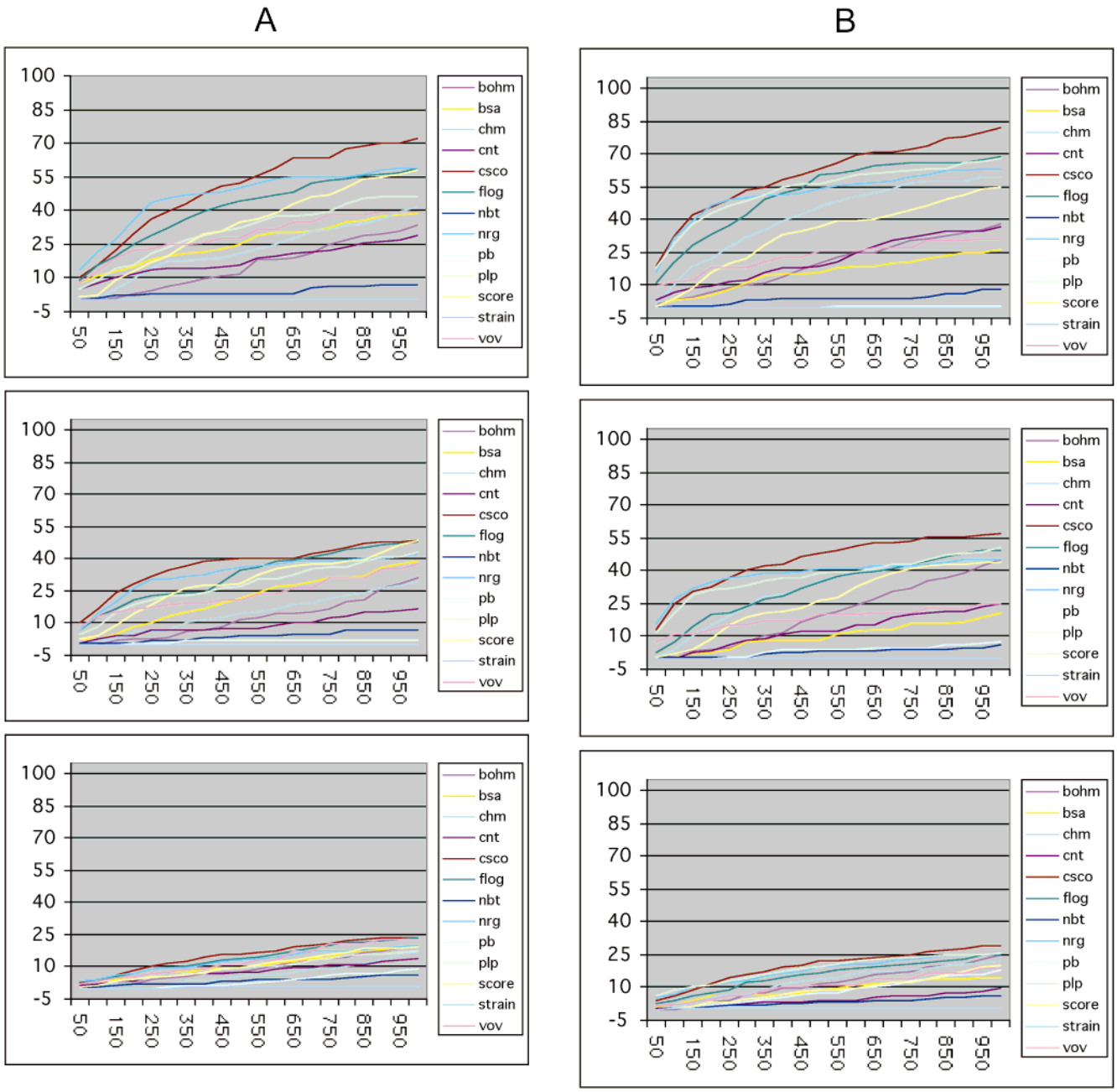
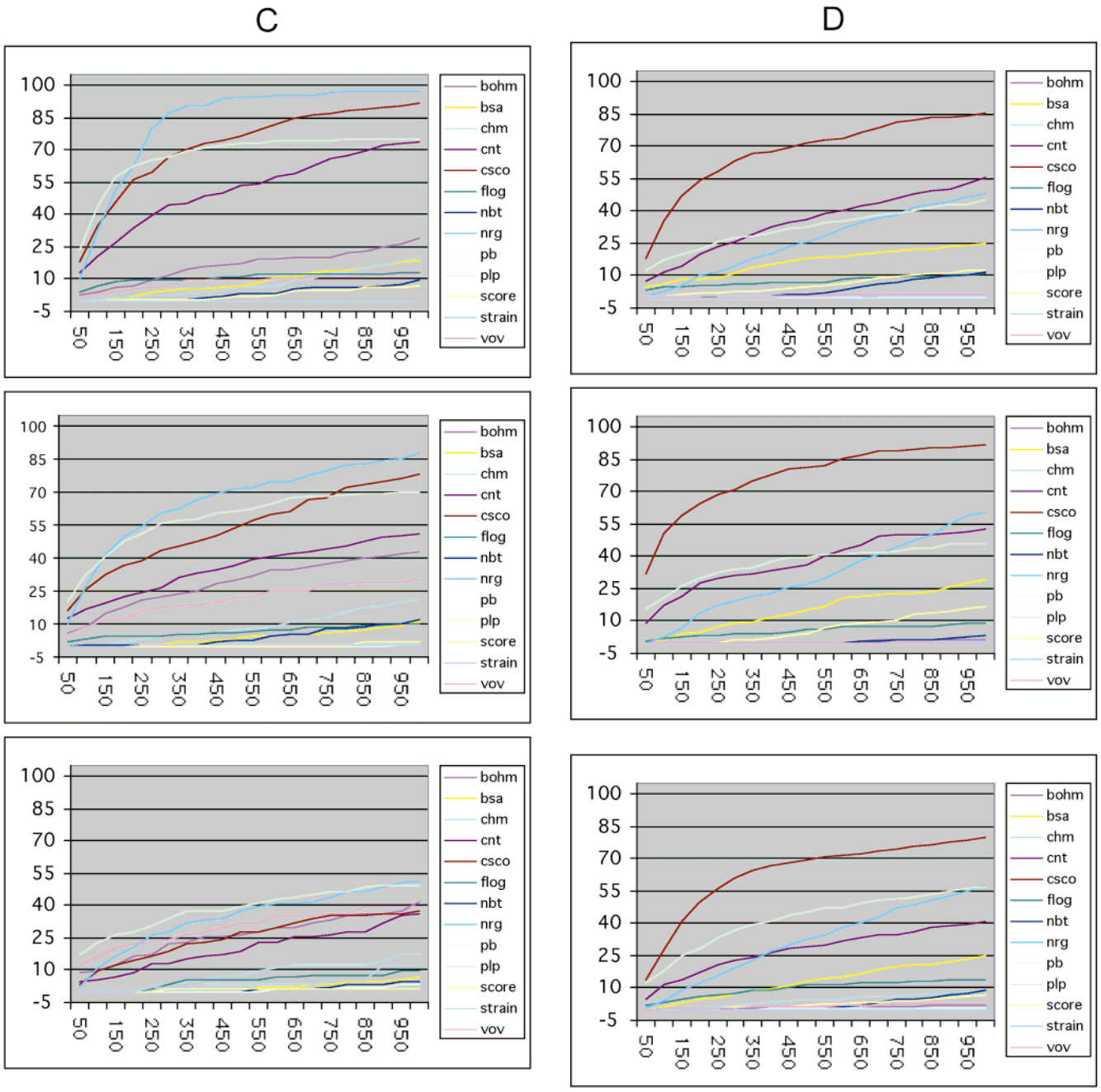


Figure 2. Performance of single scoring functions. Top panel: <100 nM. Middle panel: 100 nM − 1.0 μM. Bottom panel: 1.0 μM −30.0 μM. The *x*-axis shows the cumulative rank for the top one thousand scoring ligands considered for a given function. This includes both test compounds and randomly chosen commercial compounds. The *y*-axis gives the cumulative percentage of

training set for derivation of the ChemScore function. Additionally, HIV protease inhibitors tend to be larger molecules and even in the case of micromolar inhibitors can only bind in a limited number of acceptable binding modes. Finally, the torsion library we used to generate conformers included many examples of HIVprotease inhibitors. This allowed us to generate conformations that are very close to the "correct" conformations for these compounds.

The above analysis of single function performance identified three functions that performed well across all three targets and activity ranges: ChemScore, DOCK energy score, and PLP. Although we have exhaustively attempted all possible combinations of two, three, and four functions in a consensus approach, we will only present the results of combining the three most consistent performers denoted above. Figure 3 shows the consensus scoring results for all three test systems. In Figure 3, the *x*-axis indicates the function or combined functions of interest. The *y*-axis denotes the number of compounds common to one or more lists (functions) of the best scorers. We have chosen to present the data for depth 300 (i.e., the top 300 molecules for each function) as a conservative slice through the data. Thus, for any single function, the total list size (bar height) will be 300, but when the intersection of two or three functions is performed, the combined list size (bar height) will be less than 300 since it is unlikely that the lists will be identical. In the previous analysis of single function performance (Figure 2), it was observed that at least 50% of the active nanomolar compounds were identified in all three test cases and approximately 50% of active 100 nM compounds were correctly identified at a depth of 300. The bar graphs are separated

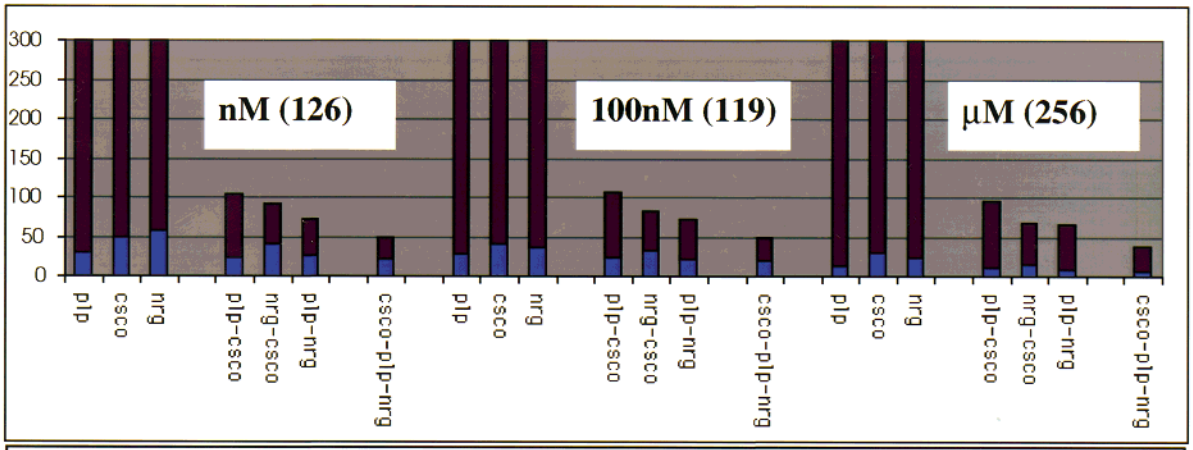


active test compounds correctly identified: $y_i = 100 \times$ (number of active test molecules retrieved among the top i scoring molecules/ total number of test molecules). A: DOCK results for p38. B: GAMBLER results for p38. C: DOCK results for IMPDH. D: DOCK results for HIV protease.

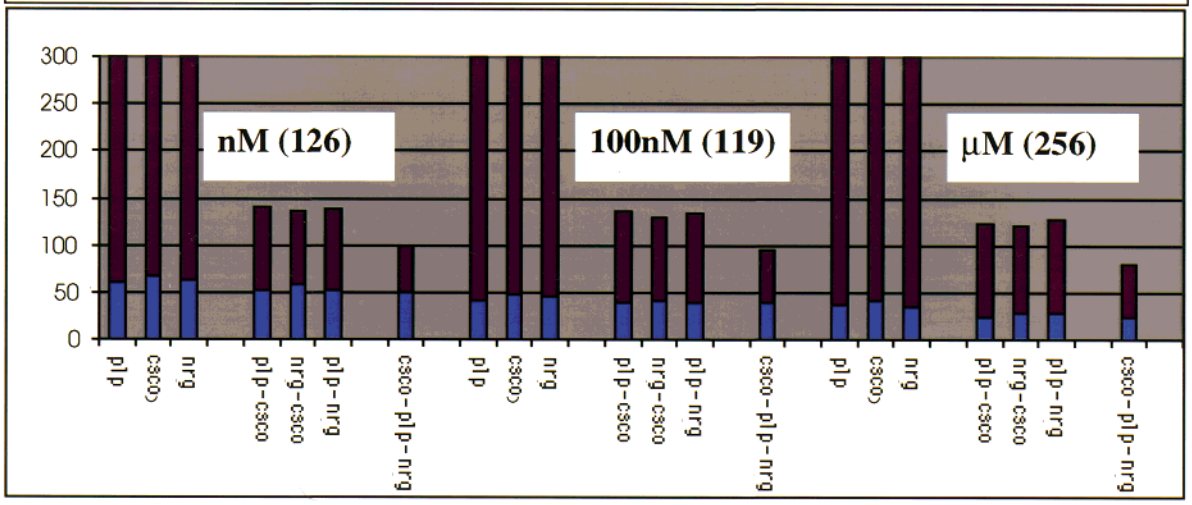
into three sections based on the three activity ranges with the labels "nM" for the compounds possessing less than 100 nM enzyme inhibitory activity; "100nM" for the compounds with activity between 100 nM and 1.0 μM; "μM" for compounds with inhibitory activity greater than 1.0 μM, but less than 30.0 μM. The total height of each bar (including both light blue and maroon sections) represents the number of total compounds (including active and inactive compounds common to the list(s)). The light blue portion of each bar is the number of active compounds common to the list(s) under consideration, while the maroon portion of the bar is the number of inactive random commercial compounds common to each list. This defines the number of false positives that would have been screened in each situation. The numbers in parentheses next to the "nM", "100nM", and "μM" labels are the number of active compounds that could have been found in that activity range. Of course, when one takes the intersection of two or more lists, the size of the intersected list will be smaller than each of the original lists unless they are identical.

Figure 3A are the results of p38 with DOCK. Recall, that p38 represents the most difficult example in this study. The most striking observation about this bar graph is the dramatic reduction in the number of false positives identified when functions are combined. This is apparent in simply proceeding from one to two functions. This graph clearly shows that without this approach one would spend a great deal of time and energy screening inactive compounds. Although this is accepted in most screening groups, it is highly inefficient especially in the context of a directed screening philoso-

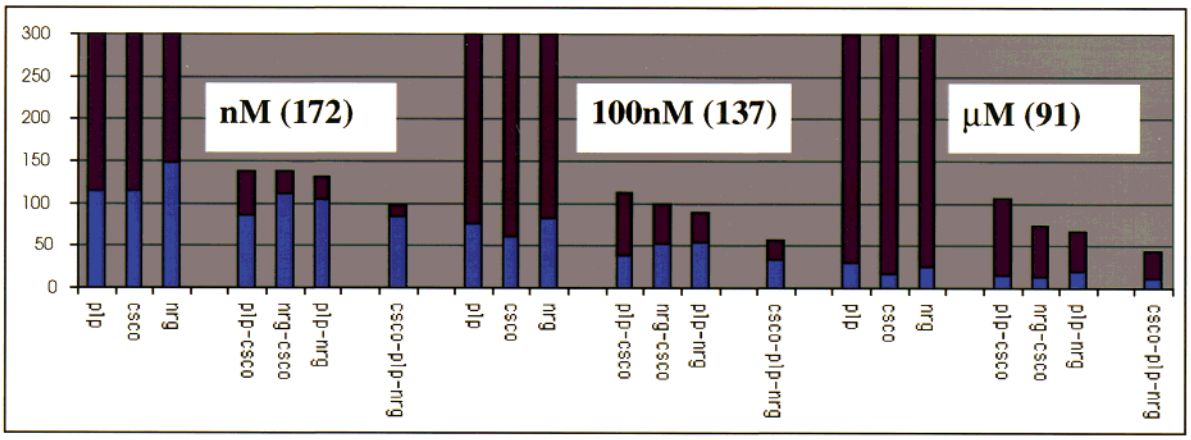
A


nM (126)
100nM (119)
μM (256)
300
250
200
150
100
50
0
plp
csco
nrg
plp-csco
nrg-csco
plp-nrg
csco-plp-nrg

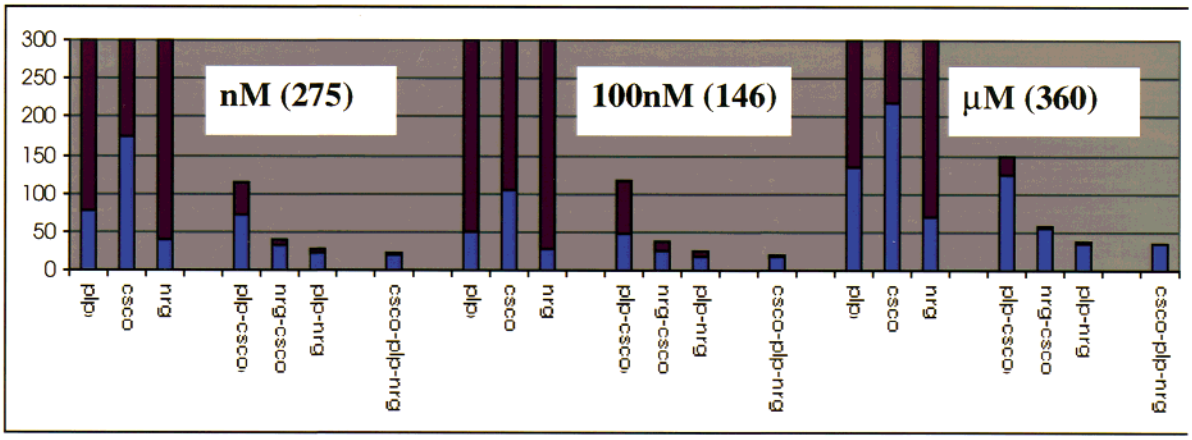
B


nM (126)
100nM (119)
μM (256)
300
250
200
150
100
50
0
plp
csco
nrg
plp-csco
nrg-csco
plp-nrg
csco-plp-nrg

C


nM (172)
100nM (137)
μM (91)
300
250
200
150
100
50
0
plp
csco
nrg
plp-csco
nrg-csco
plp-nrg
csco-plp-nrg

D


nM (275)
100nM (146)
μM (360)
300
250
200
150
100
50
0
plp
csco
nrg
plp-csco
nrg-csco
plp-nrg
csco-plp-nrg

phy. Figure 3A illustrates that combining functions not only reduces the number of false positives one would have otherwise screened, but that the total number of compounds screened could have been reduced, thus leading to an overall enrichment in the screening process. For the following discussion, we find it useful to define two terms: the "enriched hit-rate" is the number of correctly identified active compounds relative to the reduced list size produced by consensus scoring; the "raw hit-rate" would be the number of correctly identified active compounds relative to the total number of actives that could have been found in that activity range. In this instance, we would have achieved an enriched hit-rate of 18% (7 actives out of 38 total) for micromolar compounds by combining all three functions and only screening the 38 compounds common to all three lists (functions). The raw hit-rate in this case would be 3% (based on all 256 micromolar compounds). For the combinations of two functions, the enriched hit-rates would have varied from 11 to 22%, while the raw hit-rates would have varied from 3 to 6%. Of course, this approach did even better for the nanomolar and 100 nM compounds, but this is of lesser interest in practical terms.

For the case of GAMBLER applied to the same p38 problem (Figure 3B), we see a similar trend. Once again, we note the dramatic reduction in false positives in going from a single function to two or three functions. In this case, the absolute number of correctly identified actives is even higher; for micromolar compounds, we see an enriched hit rate of 29% for the combination of all three functions relative to a raw hit-rate of 9% for the same three functions. For all combinations of two functions, we see enriched hit-rates between 19 and 23% and raw hit rates between 9 and 11% for the same functions. Figure 3, C and D, show the same reduction in false positives and significant enriched hit-rates relative to raw hit-rates for IMPDH and HIV protease suggesting the generalizability of the consensus approach.

Conclusions

The most striking and useful result from this work is that consensus scoring shows much promise as a valuable method for obtaining consistent hit-rates across diverse targets, while reducing the number of false positives screened. This is imperative for our screening philosophy in which docking calculations provide only a relatively small number of compounds selected for screening purposes. The ChemScore, PLP, and DOCK energy functions have performed consistently well in this study both singly and in combination; it should come as no surprise that these functions all involved simplex minimization during the rescoring process. It has been shown [52] that simplex minimization can consistently improve both the identification of correctly docked orientations and associated improved scores. However, the lack of improvement observed when all degrees of freedom in the ligand were allowed to relax further in the post-docking MMFF minimizations suggests that even slight variations from correctly docked structures are overpenalized by the sensitivity of this method. The best performing functions in this study also share the feature of possessing relatively smooth potential surfaces with few dramatic spikes.

Some of the other functions that did not perform well in the study can be rationalized. For example, one should not expect that the Poisson−Boltzmann binding energies be very informative by themselves. Although this approach describes the Coulombic and desolvation components of binding accurately, it does not provide information on the essential hydrophobic aspects of binding. It is likely that when combined with appropriate descriptions of interaction surface shape complementarily, excluded surface area, and an estimation of conformational entropy, the PB method will have utility in the ranking of large numbers of compounds. As with MMFF, it might also be expected that the sensitivity of the PB method would overpenalize docked orientations that are close to "correct" but not perfect.

We are also not surprised that the version of the Böhm function used in this study did not perform well. This function was the first generally used empirical function and has had utility in other studies. In our experience, the lack of a repulsive term in this version was the limiting factor. We would expect that the most recently published version of this function which does account for repulsive interactions[53] would perform well on the study systems evaluated in this study. The fact that Strain energy did not perform well in this study is somewhat misleading. Since these conformers were already in or close to local minima, the inclusion of this measurement is largely uninformative. Certainly, we and others [39,54,55] have found the accurate calculation of intramolecular strain to be essential in the drug design process.

On the basis of these results, we estimate that consensus scoring as presented in this work should consistently provide hit-rates between 5 and 10% for enzymes with reasonably buried binding sites. We do not believe, however, that this approach will be generally useful for the quantitative prediction of K_i for small sets of compounds. The only potential disadvantage we can envision regarding consensus scoring is that it may not perform as well as any specific function in a specific instance, since the intersection of two nonidentical lists is, by definition, smaller than the individual lists. However, since one never knows which function might be optimal upfront, we believe the consistency and efficiency gained by consensus scoring outweighs any potential limitation. Docking errors place an upper limit on the performance of consensus scoring as they do in any computational docking/scoring experiment. The limited data we have suggests that the choice of primary scoring function (i.e., DOCK energy score for DOCK; PLP with GAMBLER) has minimal effects on the

Figure 3. Consensus scoring. The *x*-axis describes the function or functions of interest (plp = Piecewise Linear Potential, csco = ChemScore, nrg = DOCK energy score). The *y*-axis describes the number of compounds common to the top scoring compounds of each function. An arbitrary depth of 300 for the compounds common to any function or functions is shown. The bar graphs are separated into three sections based on the three activity ranges: "nM", < 100 nM; "100 nM", 100 nM − 1.0 μM; "μM", 1.0 μM −30.0 μM. The light blue portion of each bar is the number of active compounds common to the function(s) considered, the maroon portion of the bar is the number of inactive random commercial compounds common to each list or lists.

rescoring/consensus results. However, to be rigorous would require the ability to evaluate each scoring function as a primary function followed by rescoring. We have simply chosen two well-known scoring functions that have shown reliable performance in the past. It is clear that there are also other recently published scoring functions [56] that would be useful to consider in the approach presented here. Additionally, it would be interesting in future studies to examine whether the inclusion of solvation during the docking process might improve the results. Recent work[57] suggests that this might contribute to even fewer false positives than we have observed.

An issue not discussed in the present study is that of combining these same scoring functions in a more statistically rigorous manner (i.e., multiple linear regression). It is apparent that when applying the consensus scoring approach as presented here, there will be a high degree of correlation among the better performing functions. This amount of correlation is acceptable in the context of identifying micromolar leads from screening databases; however, this is not optimal in terms of deriving more quantitative and predictive models. We are presently attempting to build models that not only pick out the commonalties among the scoring functions, but also the information enriching differences between them which may enhance predictivity.[6]

We believe that given the inability of current scoring functions to capture the essential physics of ligand binding in a complete fashion, the consensus approach is a reasonable compromise. The consensus approach to scoring when combined with intelligent filtering (e.g., REOS) should provide a general improvement in docking-based selection of both commercially available compounds and combinatorial libraries in an attempt to identify and optimize leads.

Acknowledgment. The authors gratefully acknowledge the devoted work of the biophysics and enzymology groups at Vertex. We also thank the following individuals for their helpful discussions during the course of this work: Ajay, Nobuko Hamaguchi, Matt Stahl, and Anthony Nicholls. We thank all authors who made their code available to us or who supplied us with additional information necessary to reproduce their work. We thank Anthony Nicholls (OpenEye Scientific Software, Inc.) for the fast Poisson−Boltzman solver.

References

(1) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular Docking Towards Drug Discovery. *J. Mol. Recognit.* **1996**, *9*, 175−86.
(2) Jones, G.; Willett, P. Docking Small-Molecule Ligands Into Active Sites. *Curr. Opin. Biotechnol.* **1995**, *6*, 652−656.
(3) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.
(4) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(5) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeablity in Drug Discovery. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(6) Ajay On better generalization by combining two or more models: a quantitative structure−activity relationship example using neural networks. *Chemom. Intell. Lab. Styst.* **1994**, *24*, 19−30.
(7) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(8) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.
(9) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking With Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(10) Shoichet, B. K.; Kuntz, I. D. Matching Chemistry and Shape in Molecular Docking. *Prot. Eng.* **1993**, *6*, 723−732.
(11) Ewing, T. J. A.; Kuntz, I. D. Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(12) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* **1990**, *8*, 195−202.
(13) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. *J. Comput.-Aided Mol. Design* **1996**, *10*, 293−304.
(14) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated Docking of Flexible Ligands: Applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.
(15) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and An Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(16) Kramer, B.; Rarey, M.; Lengauer, T. CASP2 Experiences with Docking Flexible Ligands using FlexX. *Proteins* **1997**, *Suppl*, 221−5.
(17) Rarey, M.; Kramer, B.; Lengauer, T. Multiple Automatic Base Selection: Protein−Ligand Docking based on Incremental Construction without Manual Intervention. *J. Comput.-Aided Mol. Design* **1997**, *11*, 369−84.
(18) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−89.
(19) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−48.
(20) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor-Sites Using a Genetic Algorithm With a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(21) *FlexDock, SYBYL version 6.1*; Tripos Associates: St. Louis, MO.
(22) Charifson, P. S.; Leach, A. R.; Rusinko, A., III. The generation and Use of Large 3D Databases in Drug Discovery. *Network Sci. [Electronic Publication]* **1995**, *1*, URL: http://www.awod.com/netsci/Issues/Sept95/feature3.html.
(23) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: A Way to Enhance the Use of Molecular Docking Methods. *J. Comput.-Aided Mol. Design* **1994**, *8*, 565−82.
(24) Böhm, H.-J. The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Design* **1992**, *6*, 61−78.
(25) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical Scoring Functions. II. The Testing of an Empirical Scoring Function for the Prediction of Ligand−Receptor Binding Affinities and the Use of Bayesian Regression to Improve the Quality of the Model. *J. Comput.-Aided Mol. Design* **1998**, *12*, 503−19.
(26) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Design* **1997**, *11*, 425−45.
(27) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease − Conformationally Flexible Docking by Evolutionary Programming. *Chem. Bio.* **1995**, *2*, 317−324.
(28) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG − A System to Select Quasi-Flexible Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Design* **1994**, *8*, 153−174.
(29) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(30) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.
(31) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.
(32) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 Van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.
(33) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(34) Wilson, K. P.; McCaffrey, P. G.; Hsiao, K.; Pazhanisamy, S.; Galullo, V.; Bemis, G. W.; Fitzgibbon, M. J.; Caron, P. R.; Murcko, M. A.; Su, M. S. The Structural Basis for the Specificity of Pyridinylimidazole Inhibitors of p38 MAP Kinase. *Chem. Bio.* **1997**, *4*, 423−31.
(35) Sintchak, M. D.; Fleming, M. A.; Futer, O.; Raybuck, S. A.; Chambers, S. P.; Caron, P. R.; Murcko, M. A.; Wilson, K. P. Structure and Mechanism of Inosine Monophosphate Dehydrogenase in Complex with the Immunosuppressant Mycophenolic Acid. *Cell* **1996**, *85*, 921−30.
(36) Fleming, M. A.; Chambers, S. P.; Connelly, P. R.; Nimmesgern, E.; Fox, T.; Bruzzese, F. J.; Hoe, S. T.; Fulghum, J. R.; Livingston, D. J.; Stuver, C. M.; Sintchak, M. D.; Wilson, K. P.; Thomson, J. A. Inhibition of IMPDH by Mycophenolic Acid: Dissection of Forward and Reverse Pathways Using Capillary Electrophoresis. *Biochemistry* **1996**, *35*, 6990−7.
(37) Salituro, F. G.; Baker, C. T.; Court, J. J.; Deininger, D. D.; Kim, E. E.; Li, B.; Novak, P. M.; Rao, B. G.; Pazhanisamy, S.; Porter, M. D.; Schairer, W. C.; Tung, R. D. Design and Synthesis of Novel Conformationally Restricted HIV Protease Inhibitors. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3637−42.
(38) Baker, C. T.; Salituro, F. G.; Court, J. J.; Deininger, D. D.; Kim, E. E.; Li, B.; Novak, P. M.; Rao, B. G.; Pazhanisamy, S.; Schairer, W. C.; Tung, R. D. Design, Synthesis, and Conformational Analysis of a Novel Series of HIV Protease Inhibitors. *Bioorg. Med. Chem.* **1998**, *8*, 3631−6.
(39) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal Structure of HIV-1 Protease in Complex with VX-478, A Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181−1182.
(40) manuscript in preparation.
(41) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. MacroModel-An Integrated Software System for Modeling Organic and Bioorganic Molecules using Molecular Mechanics. *J. Comput. Chem.* **1990**, *11*, 440−467.
(42) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(43) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity: A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3222.
(44) Shah, A. V.; Walters, W. P.; Shah, R.; Dolata, D. P. Babel − A Molecular Structure Information Interchange Hub; Lysakowski, R., Gragg, C. E., Ed.; American Society for Testing and Materials: Philadelphia, PA, 1994.
(45) Clark, K. P.; Ajay Flexible Ligand Docking Without Parameter Adjustment Across 4 Ligand−Receptor Complexes. *J. Comput. Chem.* **1995**, *16*, 1210−1226.
(46) Judson, R. S.; Tan, Y. T.; Mori, E.; Melius, C.; Jaeger, E. P.; Treasurywala, A.; Mathiowetz, A. Docking Flexible Molecules − A Case-Study Of 3 Proteins. *J. Comput. Chem.* **1995**, *16*, 1405−1419.
(47) Westhead, D. R.; Clark, D. E.; Murray, C. W. A Comparison of Heuristic Search Algorithms for Molecular Docking. *J. Comput.-Aided Mol. Design* **1997**, *11*, 209−28.
(48) Wall, L.; Christiansen, T.; Schwartz, R. Programming Perl; O'Reilly: Sebastopol, CA, 1996.
(49) Honig, B.; Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **1995**, *268*, 1144−9.
(50) Nicholls, A.; Sharp, K. A.; Honig, B. Protein Folding and Association: Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons. *Proteins* **1991**, *11*, 281−96.
(51) Spencer, R. W. High-Throughput Screening of Historic Collections: Observations on File Size, Biological Targets, and File Diversity. *Biotechnol. Bioeng.* **1998**, *61*, 61−7.
(52) Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. Orientational Sampling and Rigid-Body Minimization in Molecular Docking. *Proteins* **1993**, *17*, 266−278.
(53) Böhm, H.-J. Prediction of Binding Constants of Protein Ligands: A Fast Method for the Prioritization of Hits Obtained From De Novo Design or 3D Database Search Programs. *J. Comput.-Aided Mol. Design* **1998**, *12*, 309−323.
(54) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(55) Bostrom, J.; Norrby, P.-O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Design* **1998**, 383−396.
(56) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(57) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand Solvation in Molecular Docking. *Proteins: Struct., Funct., Gent.* **1999**, *34*, 4−16.
(58) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein−Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379−394.
(59) Flower, D. R. SERF: A Program for Accessible Surface Area Calculations. *J. Mol. Graphics Modell.* **1998**, *15*, 238−244.
(60) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397.

JM990352K