Virtual Screening Using Protein−Ligand Docking: Avoiding Artificial Enrichment

Marcel L. Verdonk,* Valerio Berdini, Michael J. Hartshorn, Wijnand T. M. Mooij,
Christopher W. Murray, Richard D. Taylor, and Paul Watson

Astex Technology Ltd., 436 Cambridge Science Park, Milton Road, CB4 0QA Cambridge, United Kingdom

Received December 10, 2003

This study addresses a number of topical issues around the use of protein−ligand docking in virtual screening. We show that, for the validation of such methods, it is key to use focused libraries (containing compounds with one-dimensional properties, similar to the actives), rather than “random” or “drug-like” libraries to test the actives against. We also show that, to obtain good enrichments, the docking program needs to produce reliable binding modes. We demonstrate how pharmacophores can be used to guide the dockings and improve enrichments, and we compare the performance of three consensus-ranking protocols against ranking based on individual scoring functions. Finally, we show that protein−ligand docking can be an effective aid in the screening for weak, fragment-like binders, which has rapidly become a popular strategy for hit identification. All results presented are based on carefully constructed virtual screening experiments against four targets, using the protein−ligand docking program GOLD.

1. INTRODUCTION

In virtual screening, (large) compound libraries are screened in silico in order to increase hit rates in lead discovery. Virtual screening can be *ligand based*, i.e., using filters derived from properties of known actives, or *structure based*, where filters are based on the 3D structure of the protein. *Ligand-based* virtual screening can be done using simple one-dimensional (1D) filters (e.g. molecular weight), two-dimensional (2D) filters (e.g. substructure matching), or three-dimensional (3D) filters (e.g. 3D similarity or pharmacophore filters). *Structure-based* virtual screening can consist of 3D pharmacophore filters derived from the 3D structure of the target, but the most widely used form of *structure-based* virtual screening is protein−ligand docking.

Here, we focus on the use of protein−ligand docking, where compounds in a library are docked and scored against the 3D structure of the target. DOCK,[1] FlexX,[2] and GOLD[3,4] are probably the most widely used protein−ligand docking programs, but many more have been described in the literature.[5] Relative to ligand-based methods, protein−ligand docking is a computationally expensive virtual screening method. However, with computing power becoming cheaper, virtual screening is often done on Linux farms or PC GRIDs consisting of 10s or even 100s of processors, allowing protein−ligand docking on 10^4−10^7 compounds per week. Protein−ligand docking has already proven to be a successful virtual screening tool against several targets, including thymidylate synthase,[6] protein tyrosine phosphatase 1B[7] (ptp1b), and tRNA-guanine transglycosylase.[8] Overviews of structure-based virtual screening methods and successes were given by Lyne[9] and by Good.[10]

There are two factors that determine the performance of protein−ligand docking as a virtual screening tool. First, the docking tool needs to produce reliable binding modes (i.e. reproduce the binding modes of X-ray structures of relevant protein−ligand complexes). Although some workers have found performance to be unrelated to the quality of the binding modes,[11] we believe that structure-based virtual screening based on incorrect binding modes is meaningless and should, in carefully constructed validation experiments, perform poorly. The second performance-determining factor is the ability of the scoring function(s) used to rank the compounds. A myriad of scoring functions has been developed over the past years, and compounds can be ranked based on individual scoring functions or by consensus ranking of several scoring functions.[12,13]

Validation of virtual screening approaches is critical, particularly for such a computationally expensive tool like protein−ligand docking. These validations are typically done by comparing a set of known actives with a large number of “random” compounds. The actives are pooled with the random compounds, all compounds are docked, scored, and ranked, and the ranks of the actives are converted into enrichment plots. Such validations have been reported in the literature for various targets, scoring functions, and docking protocols. Charifson et al.[12] analyzed the performance of two different docking methods and 13 different scoring functions (and consensus combinations) against p38 MAP kinase, inosine monophosphate dehydrogenase, and HIV protease. Baxter et al.[14] validated the performance of the docking program PRO_LEADS and the *Chemscore* scoring function against thrombin, factor Xa, and the estrogen receptor. Bissantz and co-workers[11] compared the performance of GOLD, DOCK, and FlexX, and seven different scoring functions, against thymidine kinase and the estrogen receptor. The performance of the FlexX program and four scoring functions (and consensus combinations) was tested against seven targets by Stahl and Rarey.[15] Recently, Jenkins et al.[16] reported the validation of two docking protocols, DockVision/Ludi and GOLD (with their native scoring functions), against angiogenin. Good et al. described an extensive virtual screening validation exercise using GOLD, DOCK, and PROMETHEUS (a development of the PRO_LEADS program) against a range of targets.[17]

* Corresponding author phone: +44 1223 226206; fax: +44 1223 226201; e-mail: m.verdonk@astex-technology.com.

10.1021/ci034289q CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/10/2004

Recently, we tested GOLD's ability to reproduce binding modes against the CCDC/Astex validation set,[18] which contains 305 protein−ligand complexes from the Protein Data Bank[19] (PDB). At docking speeds useful for virtual screening (0.5−1 min per compound on a single CPU), and using the *Goldscore* or the *Chemscore* function to drive the dockings, GOLD reproduced the X-ray binding mode for approximately 75% of the complexes for which the ligand is "drug-like".[20] *Goldscore* is the original GOLD scoring function; *Chemscore* was originally developed by Eldridge et al.[21] and was adapted slightly for better performance with raw PDB files.

In the present study, we present the validation of GOLD as a tool for structure-based virtual screening, against four targets: neuraminidase, ptp1b, cdk2, and the estrogen receptor. The *Goldscore* function and a modified form of the *Chemscore* function are used to generate the dockings. A term that takes into account certain types of C−H···A interactions (where "A" is a hydrogen bond acceptor) was added to the *Chemscore* function; this term was added specifically to enhance performance against cdk2, as C−H···A type interactions are known to be important for kinases.[22] Three scoring functions−*Goldscore*, *Chemscore*, and an in-house version of *Drugscore*[23,24]−(and all consensus combinations) were used for scoring and ranking.

These validations are interesting in themselves; as far as we are aware, GOLD has not been validated against any of these targets before, and no validations of any docking tool have been reported against cdk2 or ptp1b. The key focus of this work, however, is to highlight a number of issues arising around the validation of structure-based virtual screening methods.

First, the choice of the random compound libraries can strongly affect the enrichments obtained. For example, if the known actives are typically much larger than the random compounds, significant enrichments can easily be obtained, but these enrichments would be meaningless; they would merely reflect a difference in 1D ligand properties (in this case molecular size) and give no indication of the usefulness of the structure-based screening approach. To avoid such biases, all our validations are run against *focused libraries*, i.e., libraries that contain compounds with 1D properties, similar to those of the actives. We illustrate the effect of the use of *focused libraries* vs *random libraries*, using simple simulations and actual virtual screening experiments.

Consensus ranking has been quite a popular method for ranking docked compounds. In this approach, docked compounds are scored with a number of scoring functions, and for each compound the different scores (or ranks) are combined into consensus scores or ranks. Consensus ranking is thought to increase hit rates, either by reducing the number of false positives[12] or by statistically reducing the errors in the scores/ranks.[25] On the other hand, in some cases, single-scoring-function ranking has been shown to outperform consensus-ranking methods.[15,26] Here, we test three protocols for consensus-ranking, as proposed by Pan et al.[25] The performance of these consensus-ranking protocols is compared systematically against that of single-scoring-function ranking.

As stated above, we believe that the ability of the docking program to produce reliable binding modes is an important prerequisite for good enrichments. The quality of the binding modes is an ill-studied parameter in virtual screening validations, presumably because often the X-ray structures are not known for the majority of the actives. In the present study, we report the performance of GOLD in terms of reproducing the binding modes of the actives for which we have X-ray structures of the complex with the target. For cdk2 and neuraminidase, we have the X-ray structures for a significant number of actives. This has allowed us to analyze the effect of the binding mode quality on the enrichments obtained.

Table 1. Overview of the Six Validation Sets Used

		X-ray			
	N	PDB	Astex	SE[a]	affinity range (M)
neuraminidase	15	15		15	10^{-10}−10^{-3}
ptp1b	25	5		25	10^{-7}−>10^{-3}
cdk2 MW<250	41	1	17	41	10^{-5}−>10^{-3}
cdk2 MW>250	23	11	6	23	10^{-8}−10^{-4}
ER agonists	20	3		3	10^{-10}−10^{-7}
ER antagonists	17	2		2	10^{-10}−10^{-7}

[a] Number of actives for which there is structural evidence, either from X-ray crystallography or from direct binding methods, that they bind in the binding site studied.

It is often known that actives form certain interactions with the target, e.g. a specific hydrogen bond. In such cases, this additional structural knowledge, in the form of a pharmacophore, can be used in structure-based virtual screening to improve the binding modes produced and/or to improve the enrichments. Several examples have been reported in the literature where such pharmacophores were used as part of the docking protocol. Fradera et al.[27] combined the DOCK scoring function with a function representing the overlap between the docked ligand and that of the X-ray structure of a known active, in complex with the target. Recently, Hindle et al.[28] incorporated docking under pharmacophore constraints in FlexX and showed increases in enrichments against thermolysin, carbonic anhydrase, and dihydrofolate reductase. Good et al. also showed the advantages of using pharmacophores in virtual screening. Here, we implement pharamacophoric restraints in GOLD and investigate their effect on the enrichments for low-molecular-weight cdk2 binders and for neuraminidase actives.

Finally, in recent years, screening for low-molecular-weight, weakly-binding compounds has become a popular strategy for hit identification.[8,29−34] But, as far as we are aware, no validation of structure-based virtual screening methods on test sets of fragment-like actives has been reported in the literature. We have created a separate, low-molecular-weight cdk2 test set, for which the molecular weights of all actives are below 250; the affinities of these compounds range from mid-μM to weaker than mM. Also, half of the actives in the ptp1b test set have molecular weights below 250 and similarly low affinities.

2. MATERIALS AND METHODS

Target Structures. Table 1 lists the numbers of actives in our test sets for which X-ray structures of the complex with the target are available in the PDB or in the Astex in-house structure database. For each target, these structures were superimposed based on the residues in the active site. In the resulting frame of reference, the ligands (henceforth known as the "reference ligands") were saved separately from the proteins. Hydrogen atoms were added to the protein structures used in the virtual screens, ensuring that protonation states are correct; all water molecules were removed from the structures. The binding sites were defined using the reference ligands; all protein atoms within 6 Å of a non-hydrogen atom in any of the reference ligands were included. For each target, dockings were performed against a variety of structures, and the structures that are most promiscuous, i.e., against which most compounds are docked correctly, were used in the virtual screens.

In our experience, it is difficult to dock estrogen receptor agonists correctly against the antagonist form of the receptor and vice versa. For this reason, and because the antagonists are considerably larger than the agonists (see below), we decided to construct two separate validation sets. The agonists were docked against PDB structure 1qkm, the closed form of the receptor, and a complex with the agonist genistein; the antagonists were docked against PDB structure 1qkn, the open form of the receptor, and a complex with the antagonist raloxifene.

For ptp1b, we used PDB entry 1pty, a complex with phosphotyrosine. This structure corresponds to the closed form of the enzyme (WPD loop is closed), which is the conformation typically observed when drug-like compounds are bound to ptp1b.

In cdk2, the side chain conformations of Lys33 and Lys89 can vary significantly between structures. In addition, minor main-chain and side-chain movements can occur in the vicinity of His84. These protein movements clearly affect the steric properties of the binding site and appear to be, at least partially, ligand-induced. For example, if we take a structure from a cdk2 complex with a fragment-like binder, larger cdk2 actives do not sterically fit into the binding site. We could (to some extent) account for this ligand-induced effect by docking against two separate cdk2 structures, one for the larger compounds, and one for the smaller compounds. Hence, we split the cdk2 inhibitors in two groups, one containing the higher-molecular-weight binders and the other the lower-molecular-weight ones (see below). The larger compounds were docked against PDB structure 1di8, for which the binding pocket is relatively “open” in the Glu81-Leu83 region; the smaller compounds were docked against an in-house structure, p934, which has a small fragment-like ligand bound (see Figure 1)[35] and for which the binding site is much narrower in the Glu81-Leu83 region.

For neuraminidase, we used PDB 1mwe. This structure represents the “closed” form of the enzyme, and in a previous study, it proved to be the most promiscuous, i.e., most actives were docked correctly.[36]

Sets of Known Actives. Table 1 lists the sources of the known actives in our test set; they originate from both literature and from our in-house compound collection. Each set of compounds typically spans a range of different compound classes, although we have not insisted each compound represents a separate chemotype (see ref 17). For both cdk2 and the estrogen receptor, we decided to split the known actives into two separate sets. This was done for two principal reasons. In both cases, the distribution of the heavy-atom-counts for the actives (see Figure 2) is bimodal, which is not ideal (see below). Also, in both cases, the target undergoes a ligand-induced fit, which affects the ability of GOLD, and any other docking program, to predict the binding mode correctly (see above). Therefore, the estrogen receptor actives were split into agonists and antagonists; the cdk2 actives were split into the lighter (MW < 250) and heavier compounds (MW > 250). The set of low-molecular-weight cdk2 actives is particularly challenging: It contains *only* weak to moderate affinity fragment-like binders.

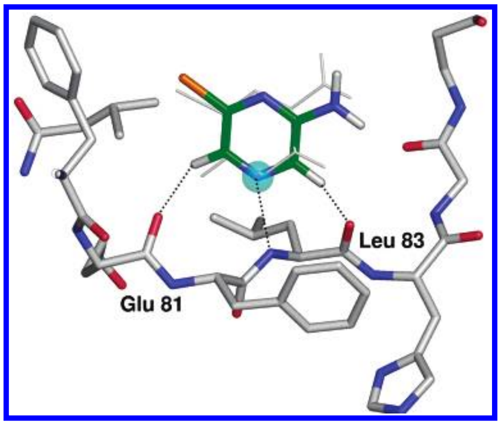

Figure 1. X-ray structure of AT934 in complex with cdk2.[35] The pharmacophore acceptor point used to guide the dockings is shown in cyan. In gray, a *Chemscore* docking (produced without the pharmacophore restraint) of AT934 is shown. Without the pharmacophore, this binding mode is not always found (although it *is* the global optimum); with the pharmacophore, the experimental binding mode is found five out of five times.

We attempted to bias our sets of actives to compounds for which we have structural evidence for binding or ideally, for which we know the binding mode. For five of the 37 estrogen receptor actives, an X-ray structure of the complex with the receptor is available in the PDB. The same is true for ptp1b, i.e., we have the structures of five of the 25 actives, complexed with ptp1b. The remaining 20 ptp1b actives are from the Astex in-house compound collection. For all these compounds, we have confirmed, by means of a direct binding method, that they are displaced by “compound 5” from Iversen et al.,[37] and hence are likely to bind in the phosphate-binding pocket. Structures are available in the PDB for all 15 neuraminidase actives; we have ensured that, in these structures, there are no protein−ligand clashes, crystallographic contacts, or unlikely ligand geometries.[36] For 12 of the cdk2 actives, structures are available in the PDB. For the remaining 52 actives, we confirmed, by X-ray crystallography, that they bind in the ATP-binding pocket; for 23 of these actives, the electron density allowed a reliable determination of the binding mode.

Library Design. We, like many other workers, test the performance of our virtual screening applications by comparing the results (scores) obtained for the active compounds, with those obtained for a library of compounds that are assumed to be inactive. The active compounds are pooled together with a large set of inactive compounds, and then all compounds are docked, scored, and ranked. The fraction of active compounds ranked near the top of the list determines the performance of the application. Here, we pick our inactive compounds from an in-house database, ATLAS, which contains approximately 3.1 million compounds from High-Throughput Screening suppliers. The critical factor, however, is how the inactives are selected from the ATLAS database.

The tool we use here for virtual screening is protein−ligand docking, which involves the scoring of the interactions

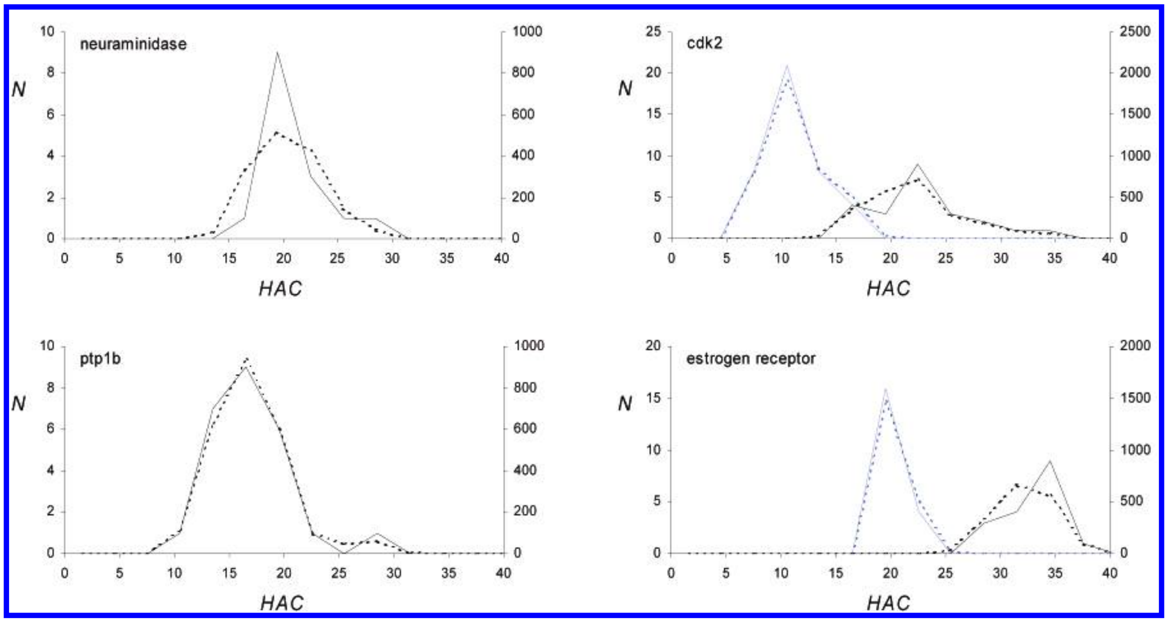

Figure 2. Heavy-atom-count distribution for the six sets of actives (solid lines) and for the *focused libraries* (dotted lines). The distributions for the low-molecular-weight cdk2 compounds and the estrogen receptor agonists are shown in blue.

formed between protein and ligand. To assess the performance of this three-dimensional scoring technique, we need to eliminate, as much as possible, any biases in lower dimensions. Assume, for example, that the actives were all sugar-like compounds, with four or five OH groups, and the inactive compounds were selected, truly randomly, from ATLAS. On average, compounds in ATLAS have fewer than two OH groups. Therefore, any scoring function that rewards OH groups in ligands would yield good enrichments. The question here, however, is if docking against the 3D structure of the target provides any benefits over simply counting the OH groups in the ligands. Perhaps the most important one-dimensional parameter in this context is that of molecular size. Larger compounds typically give higher scores, because they have a larger surface area to interact with the protein (this is a general limitation of empirical scoring functions). Hence, a large number of unspecific interactions formed by a larger compound can easily outweigh a small number of highly specific interactions formed by a smaller compound.

Figure 3 illustrates how the molecular size distribution in the library of inactive compounds can cause artificial enrichments or even negative enrichments. Figure 3a shows a typical plot for the dependency of the *Chemscore* function (eq 10) on the size of the compounds. Figure 3b shows the heavy-atom-count distributions for the compounds in ATLAS that pass Lipinski's "Rule of Five",[38,39] for a set of actives with relatively low molecular weight, and for a set of actives with relatively high molecular weight. If the ATLAS set (or a random subset) is used as the library of inactive compounds, the low-molecular-weight set of actives has a size disadvantage compared to the inactives, and the high-molecular-weight set has a size advantage. We assume that the scores of the actives have the same dependency of molecular size as the library compounds (Figure 3a, i.e., no real enrichment). In that case, the simulated enrichment graphs in Figure 3c, generated by combining parts a and b of Figure 3, show that virtual, positive enrichments are obtained for the high-molecular-weight set of actives; for the low-molecular-weight set, negative enrichments (i.e. worse than random) are obtained. In both cases, the virtual enrichments are caused purely by the difference in the size of actives and inactives.

To prevent such virtual enrichments, all validations presented here are done against *focused libraries*, containing compounds with 1D properties similar to those of the active compounds. Here, we used three simple 1D properties, roughly describing the physical nature of the compounds: (i) number of hydrogen-bond donors, N_D; (ii) number of hydrogen-bond acceptors, N_A; (iii) number of nonpolar atoms, N_{NP}. For each set of actives, a separate *focused library* was generated as follows. The "distance", D, between two actives, i and j, is defined as

$$D(i,j) = \sqrt{(N_D(i) - N_D(j))^2 + (N_A(i) - N_A(j))^2 + (N_{NP}(i) - N_{NP}(j))^2} \quad (1)$$

First, for each active, i, we calculate the distance to the nearest other active, *D(i)*. *D(i)* is then averaged over all actives to give the average distance to the nearest other active, D_{min}. Next, for each active in the validation set, 100 compounds are selected at random from ATLAS, ensuring that for each selected compound the distance to the active is less than D_{min}. Additionally, the compounds selected from ATLAS can only contain the elements C, H, O, N, S, P, F, Cl, Br, I. As a result, the distributions of the 1D properties of the compounds in the *focused library* are similar to those of the known binders. Figure 2 shows the heavy atom count distributions of the six sets of actives and of the *focused libraries* generated using the approach outlined above.

Docking. GOLD was used to dock and score all compounds against the target binding sites. We used the previously described version of GOLD,[20] with two minor alterations: (i) The torsion library was extended to include a wider range of rotatable bonds. (ii) In addition to CH carbons, F, Cl, Br, I, and nonpolar sulfur atoms in the ligands are mapped onto the "hydrophobic fitting points" (together with the hydrogen-bonding fitting points, these points are used to place the ligand in the binding site[4]).

GOLD uses a Genetic Algorithm (GA) to explore the possible binding modes. The GA modifies all dihedrals of

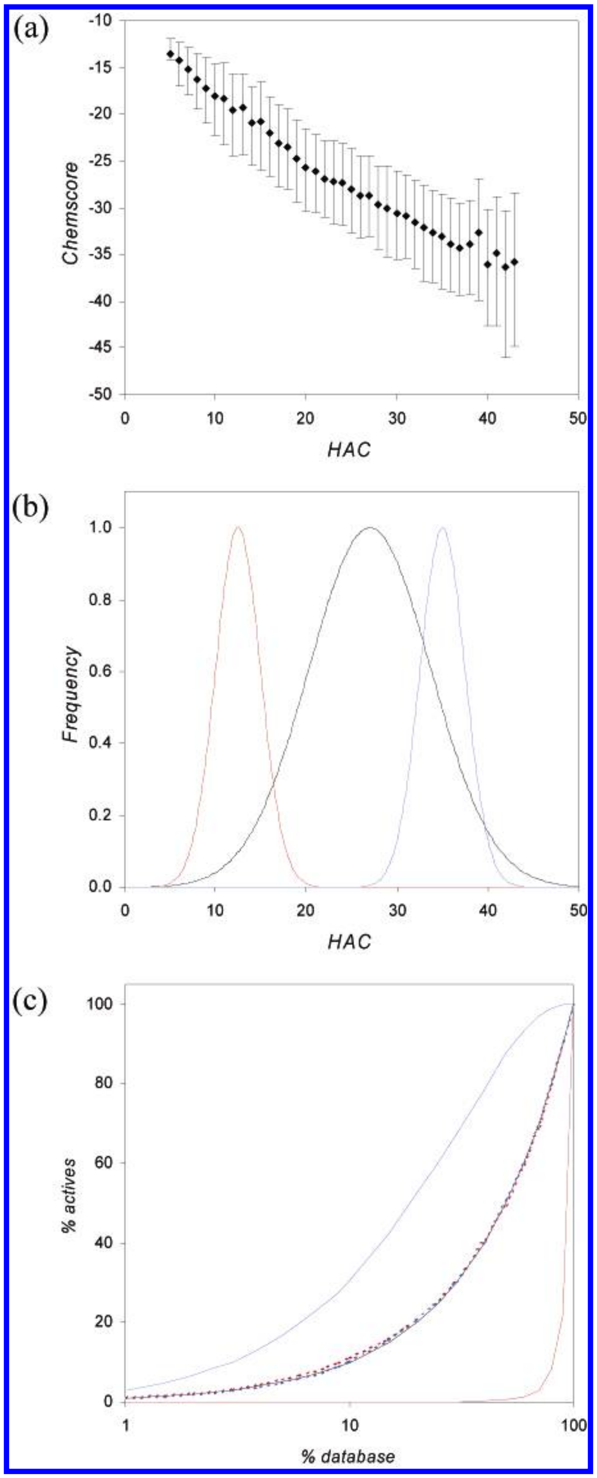

Figure 3. Simulation of the effect of molecular size on enrichment rates. (a) A typical dependency of the *Chemscore* on the heavy-atom count; it is a result of the docking of about 100 000 random compounds, with varying molecular weights, against the cdk2 binding site; the error bars reflect the variance in the *Chemscore* values; (b) heavy-atom-count distributions of two hypothetical sets of actives—one low-molecular-weight (blue), the other higher molecular weight (red)—and for a drug-like library (a subset of compounds from ATLAS that pass Lipinski's Rule of Five); (c) simulated enrichment graphs when the two sets of actives are screened against the drug-like library (solid lines), assuming that the scores of both actives and library compounds have the dependency of molecular size that is shown in 3 part a; enrichment plots are also shown when the two sets of actives are screened against focused libraries (dotted lines).

ligand rotatable bonds, ligand ring geometries (by flipping ring corners), dihedrals of protein OH groups and NH_3^+ groups, and the placement of the ligand in the binding site. We used the *GOLD Default 4* GA settings,[20] except where indicated differently. Hence, for each compound, GOLD performs 10 dockings, each consisting of 10 000 GA operations. GOLD terminates early when the top three dockings are within 1.5 Å of each other. Each docking is followed by a Simplex optimization in which all ligand (and protein OH/NH_3^+) torsions and the position and orientation of the ligand are refined to the nearest local optimum.

Recently, we restructured GOLD, such that the implementation of different scoring functions is now straightforward.[20] Here, we investigated three different scoring functions to drive the protein–ligand docking, *Goldscore*, *Chemscore*, and *Drugscore*.

Goldscore, the original GOLD scoring function, is a molecular-mechanics-like function, and the function that is optimized during docking has four terms

$$\mathit{GS\ Fitness} = S_{\mathrm{hb\_ext}} + S_{\mathrm{vdw\_ext}} + S_{\mathrm{hb\_int}} + S_{\mathrm{vdw\_int}} \quad (2)$$

where $S_{\mathrm{hb\_ext}}$ is the protein–ligand hydrogen-bond score and $S_{\mathrm{vdw\_ext}}$ is the protein–ligand van der Waals score. $S_{\mathrm{hb\_int}}$ is the contribution to the *GS Fitness* due to intramolecular hydrogen bonds in the ligand; this term is switched off in all calculations presented in this work; $S_{\mathrm{vdw\_int}}$ is the contribution due to the intramolecular strain in the ligand.

The *Chemscore* function was originally developed by Baxter et al.[20,21,40] Recently, we implemented *Chemscore* as a scoring function for GOLD, and, to improve performance with "raw" PDB files, we made several alterations to the functional form and parameters.[20] Here, two additional alterations were made to the *Chemscore* function: (i) A term was added to score CH···A type interactions (where A is an acceptor); this term was introduced especially for kinases, for which these types of interactions have been shown to be important;[22] this has hardly any effect on the results obtained for the other targets presented here, but for cdk2, including the CH···A term gives significant improvements in both the docking accuracies and enrichments. (ii) In order to prevent certain unrealistic ligand conformations, the internal energy term for the ligand was modified, such that hydrogen–hydrogen clashes are taken into account. Hence, the functional form of the *Chemscore* function we used in this work to drive the dockings is

$$\mathit{CS\ Fitness} = \Delta G_{\mathrm{o}} + \Delta G_{\mathrm{hbond}} S_{\mathrm{hbond}} + \Delta G_{\mathrm{cha}} S_{\mathrm{cha}} + \Delta G_{\mathrm{metal}} S_{\mathrm{metal}} + \Delta G_{\mathrm{lipo}} S_{\mathrm{lipo}} + \Delta G_{\mathrm{rot}} H_{\mathrm{rot}} + E_{\mathrm{clash}} + E_{\mathrm{int}} \quad (3)$$

where S_{hbond}, S_{metal}, and S_{lipo} are scores for hydrogen-bonding, acceptor-metal, and lipophilic interactions, respectively; H_{rot} is a score representing the loss of conformational entropy of the ligand upon binding to the protein; E_{clash} is the protein–ligand clash energy, and E_{int} is the ligand internal energy; the ΔG terms (except ΔG_{cha}) are coefficients derived from a multiple linear regression analysis on a training set of 82 protein–ligand complexes from the PDB.[21] The functional form of all these terms was described before.[21,40]

The score for CH···A type interactions, S_{cha}, has the same functional form as the hydrogen-bond score, i.e.:

$$S_{\mathrm{cha}} = \sum_{\mathrm{CA}} f(\Delta r_{\mathrm{CA}},\Delta r_1,\Delta r_2) f(\Delta\alpha_{\mathrm{CA}},\Delta\alpha_1,\Delta\alpha_2) f(\Delta\beta_{\mathrm{CA}},\Delta\beta_1,\Delta\beta_2) \quad (4)$$

with $\Delta r_{\mathrm{CA}} = |r_{\mathrm{CA}} - r_{\mathrm{o}}|$, $\Delta\alpha_{\mathrm{CA}} = |\alpha_{\mathrm{CA}} - \alpha_{\mathrm{o}}|$ and $\Delta\beta_{\mathrm{CA}} = |\beta_{\mathrm{CA}} - \beta_{\mathrm{o}}|$, and with

$$f(x,x_1,x_2) = \begin{cases} 1 & x \le x_1 \\ (x_2 - x)/(x_2 - x_1) & x_1 < x \le x_2 \\ 0 & x > x_2 \end{cases} \quad (5)$$

where r_{CA} is the H···A distance, α_{CA} is the C−H···A angle, and β_{CA} is the R−A···H(C) angle for a given CH-acceptor pair. r_{o}, α_{o}, and β_{o} are the ideal values for r_{CA}, α_{CA}, and β_{CA}, respectively (we used $r_{\mathrm{o}} = 2.35$ Å, $\alpha_{\mathrm{o}} = 180°$, and $\beta_{\mathrm{o}} = 180°$). Δr_1, Δr_2, $\Delta\alpha_1$, $\Delta\alpha_2$, $\Delta\beta_1$, and $\Delta\beta_2$ are constants that control the deviation from the ideal CH···A geometry (we used $\Delta r_1 = 0.25$ Å, $\Delta r_2 = 0.65$ Å, $\Delta\alpha_1 = 50°$, $\Delta\alpha_2 = 100°$, $\Delta\beta_1 = 70°$, and $\Delta\beta_2 = 80°$). In this study, only CH groups in aromatic rings were considered, where the carbon atom is adjacent to a ring nitrogen atom (these types of CH groups are most often seen to form C−H···A type interactions). The summation in eq 4 is over all combinations of CH groups and acceptors in protein and ligand.

The functional form of the internal clash energy of the ligand remained unchanged. The list of ligand atom pairs that contribute to the clash energy was simply extended to include all ligand hydrogen−hydrogen pairs, for which the two hydrogens are separated by at least four bonds. We used $r_{\mathrm{clash}} = 2.0$ Å for the hydrogen−hydrogen clash distance.

A *Drugscore*-type scoring function was implemented in GOLD, following the work of Gohlke et al.[23,24] We did not include the solvent-accessible surface-dependent singlet potential, but we did add a term to account for the ligand internal energy. Hence, the functional form of the *Drugscore*-like function that was used to drive the dockings is

$$DS\ Fitness = \sum_p \sum_l \Delta W_{\mathrm{pl}}(r_{\mathrm{pl}}) + C_{\mathrm{i}} \cdot E_{\mathrm{int}} \quad (6)$$

where E_{int} is the *Chemscore* ligand internal energy, and C_{i} is a scale factor (after some optimization, we used $C_i = 1.0$). The summation is over all combinations of protein atoms, p, and ligand atoms, l, and r_{pl} is the distance between protein atom p and ligand atom l. $\Delta W_{\mathrm{pl}}(r)$ is the distance-dependent pair-potential for protein atom p and ligand atom l, as defined by Gohlke et al.[23,24] To speed up scoring and docking, grids were precalculated for each atom type.

To derive the pair-potentials, a database of protein−ligand complexes was constructed from the PDB. Ligand bond types were assigned using an in-house program,[41] which uses a combination of rules and algorithms from BALI[42] and an approach developed by Sayle.[43] Bond types for proteins were assigned based on residue and atom names. Each ligand was classified as *normal*, *covalent*, or *cofactor*. The assignment of *cofactors* was based on the following common residue names: HEM, NAD, FAD, ADP, ATP, NAP, NDP, GDP, and IDP. *Covalent* ligands were assigned based on short contacts between protein and ligand. Only binding sites for *normal* ligands were added to the database. Crystallographic symmetry was used to expand the binding sites if needed; symmetry-related protein and ligand atoms were generated up to a radius of 15 Å around each ligand. If a ligand occurs more than once in a single PDB entry, and the binding sites are (pseudo) symmetric, only the first occurrence of the ligand is used. Pair potentials were generated for all atom types used by Gohlke et al.[23]

Table 2. Success Rates[a] for Binding Mode Predictions against the CCDC/Astex Validation Set,[18] Using *Goldscore*, *Chemscore*, and *Drugscore*[b]

	N	*Goldscore*	*Chemscore*	*Drugscore*
clean list[c]	224	63.0(1.8)	60.9(1.8)	51.6(1.3)
drug-like list[d]	139	73.4(2.1)	74.8(2.1)	65.9(2.3)
fragment-like list[e]	79	76.4(2.8)	78.6(3.9)	67.8(3.6)

[a] Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode, using the *GOLD Default 4* GA settings.[20] [b] All values are averages over 50 runs; standard deviations are given in parentheses. [c] The "clean list" is a subset of the CCDC/Astex validation set for which complexes do not exhibit protein−ligand clashes, crystallographic contacts, or unlikely ligand geometries; for closely related complexes, only one representative was kept.[18] [d] A subset of the clean list, for which ligands have 10 or fewer rotatable bonds and a polar surface area equal to or less than 140 Å^2.[20,44] [e] A subset of the clean list for which ligands are not covalently bound to the protein and have 5 or fewer rotatable bonds and between 7 and 20 non-hydrogen atoms.[20]

Table 2 shows the docking success rates of the three functions, against the CCDC/Astex validation set.[18] Compared to the *Goldscore* and *Chemscore* functions, the performance of the *Drugscore* function, in terms of producing reliable binding modes, is poor. These results are roughly in line with those obtained by Sotriffer et al.[45] when they used the *Drugscore* function for protein−ligand docking. Hence, we decided not to use the *Drugscore* function to generate the binding modes but only to score the compounds (see below).

Pharmacophore Restraints. Although we believe the empirical scoring functions used here for protein−ligand docking are state-of-the-art, they are far from perfect. In some cases it will be beneficial to use structural knowledge about the way inhibitors typically bind to a given target, to guide the docking. Here we do that by adding a pharmacophore term to the scoring function, i.e.:

$$Fitness' = Fitness + C_{\mathrm{p}} \cdot Pharmscore \quad (7)$$

This functionality was added to all three scoring functions described here. C_{p} is a scaling factor to adjust the weight of the pharmacophore relative to the scoring function. *Pharmscore* is defined as follows

$$Pharmscore = \sum_i c_i p_i \quad (8)$$

where the summation is over all pharmacophore points in the pharmacophore. *Pharmacophore* points can be of the following types:

donor: matches hydrogen-bond donors

acceptor: matches hydrogen-bond acceptors

donor/acceptor: matches atoms that are both donor and acceptor

hydrophobic: matches hydrophobic atoms (using the *Chemscore* definition)

any: matches any non-hydrogen atom

rule: a SMARTS pattern[46] defines the atom types matched.

p_i is the best overlap of a matching ligand atom with pharmacophore point i, making sure that each ligand atom only contributes once to *Pharmscore*. c_i is the normalized weight of pharmacophore point i.

Two shapes for the overlap function p were implemented: *block shaped* and *Gaussian*, both have a cutoff distance, d_o. The *block shaped* function has a value of 1.0 when a pharmacophore point is within d_o of a matching ligand atom, and 0.0 when it is outside. For the Gaussian shaped overlap function, $p = \exp(- [d/d_o]^2)$, where d is the actual distance between pharmacophore point and matching ligand atom.

Single Scoring Function Ranking. All three scoring functions described above were used to score and rank the compounds. Dockings produced with the *Goldscore* function were rescored with the *Chemscore* and *Drugscore* functions; dockings produced using the *Chemscore* function were rescored with the *Goldscore* and *Drugscore* functions. For each rescore, the dockings are reoptimized in the second scoring function, using the Simplex algorithm. This leads to six single-scoring-function ranking protocols: G–G, G–C, G–D, C–G, C–C, and C–D (where the first character indicates the scoring function used for docking ("G" for *Goldscore*, "C" for *Chemscore*, and "D" for *Drugscore*), and the second character is the scoring function used for scoring and ranking.

At the scoring and ranking stage, we use modified versions of the expressions for the *Goldscore*, *Chemscore*, and *Drugscore* functions given in eqs 2, 3, and 6, respectively. The problem with the functions used to drive the dockings is that they contain ligand internal energy terms, which have an arbitrary reference point and, although essential for successful docking, are meaningless when different compounds are compared. However, in theory, the intramolecular energy could be of some use if it could be calculated relative to that of the free ligand. We attempted to do that here by keeping track—during the docking—of the best internal energy of the compound and subtracting that from the *Fitness* function; note that this is only possible for the scoring function that is used to drive the docking. For the *Chemscore* function, this had a significant positive effect on the enrichments obtained. For *Goldscore*, on the other hand, it turned out to have a detrimental effect on the enrichments. We have noticed that, in certain cases, the internal energy term of the *Goldscore* function can be very sensitive to relatively small variations of the ligand conformation, which could possibly explain the observed effect. Hence, the functional forms of the scores that were used to rank compounds and calculate enrichments are

$$\mathit{Goldscore} = \mathit{GS\ Fitness} - S_{\mathrm{hb\_int}} - S_{\mathrm{vdw\_int}} \quad (9)$$

$$\mathit{Chemscore} = \mathit{CS\ Fitness} - E_{\mathrm{int}}(\mathit{free}) \quad (10)$$

$$\mathit{Drugscore} = \mathit{DS\ Fitness} - C_i \cdot E_{\mathrm{int}} \quad (11)$$

where $E_{\mathrm{int}}(\mathit{free})$ is the best observed internal energy of the free ligand, during the course of a *Chemscore* docking; $C_{\mathrm{i}} = 1.0$; note that, when the *Chemscore* function is used to rescore *Goldscore* dockings, $E_{\mathrm{int}}(\mathit{free}) = E_{\mathrm{int}}$.

Consensus Ranking. In consensus ranking, multiple scoring functions are combined to rank compounds. In analogy with Wang and Wang,[47] we investigated three possible strategies for consensus ranking: (i) *rank-by-number*, where the scores are averaged (after normalizing) and compounds reranked based on their average score; (ii) *rank-by-rank*, where the compounds are first ranked in each individual scoring function, the ranks are averaged, and compounds are reranked based on their average rank; (iii) *rank-by-vote*, where compounds have to rank in the top x% of the ranked lists of each of the individual scoring-functions-combined.

The *rank-by-rank* method is straightforward. The *rank-by-number* method, however, has the complication that the scores combined are typically not on the same scale. Here, we brought the three scoring functions on the same scale by converting them to binding energy estimates. To do this, we used 60 complexes from the CCDC/Astex validation set, for which we have reliable affinity data. For these complexes, we carried out a linear regression of the experimental binding energies against each of the three scoring functions. This resulted in the following expressions for the binding energy estimates, from each of the three scoring functions

$$\Delta G_{\mathrm{binding}}(GS) = -\,0.450 \cdot \mathit{Goldscore} - 9.489 \qquad R^2 = 0.55,\ s = 9.3\ \mathrm{kJ/mol} \quad (12)$$

$$\Delta G_{\mathrm{binding}}(CS) = 0.911 \cdot \mathit{Chemscore} - 8.190 \qquad R^2 = 0.53,\ s = 9.6\ \mathrm{kJ/mol} \quad (13)$$

$$\Delta G_{\mathrm{bindin}g}(DS) = 0.209 \cdot \mathit{Drugscore} - 23.071 \qquad R^2 = 0.29,\ s = 11.7\ \mathrm{kJ/mol} \quad (14)$$

where R is the correlation coefficient and s is the standard error in the predicted free energies of binding. For the *rank-by-number* consensus ranking, these three estimates were averaged, and compounds were reranked based on this average score. For *Drugscore*, the correlation with affinity is poor, but the R^2 value is roughly in line with what was observed by Gohlke et al.[24] for a diverse set of 71 complexes ($R^2 = 0.33$). Although the scoring functions used here are state-of-the-art, the correlation with experimental binding affinity is relatively poor (a general observation for fast scoring functions). This is not necessarily disastrous for virtual screening applications though, because there the objective is to identify potential binders in a database of mainly inactive compounds, rather than ranking a set of known binders.

In our implementation of the *rank-by-vote* approach, compounds must be present in the top x% of the ranked lists of each of the individual scoring-functions-combined. The complication with the *rank-by-vote* approach is that it is difficult to control the number of compounds that pass the consensus test. For example, if we combine two scoring functions, and use $x = 3\%$, the fraction of compounds that pass the consensus test, could be anything between 0 and 3%. This makes it difficult to compare enrichment rates obtained using this consensus approach with other single-scoring-function or consensus ranking techniques. We solved this problem by introducing the desired fraction of compounds, y. Next, x, starting at $x = y$, is increased until the fraction of compounds that pass the consensus test equals y.

We tested all combinations of the three single scoring functions, *Goldscore*, *Chemscore* and *Drugscore*; together with the two docking protocols, this results in eight additional consensus-ranking protocols: G−GC, C−GC, G−GD, C−GD, G−CD, C−CD, G−GCD, and C−GCD (using the same abbreviations as above).

Virtual Screening. All virtual screens were performed using a Web-based, in-house virtual screening platform.[48] The compounds in ATLAS are stored as SMILES strings.[49] For each of the six validation sets, *focused compound libraries* were prepared from ATLAS, using the approach described above. SMILES strings were prepared manually for all the known actives, resulting in six small *active compound libraries*. Next, each library was virtually screened against the corresponding target, using the following protocol: (i) compound SMILES strings are charged using a fixed set of rules; (ii) compound 3D input structures are generated from the SMILES strings using Corina;[50] (iii) compounds are docked against the target using GOLD and either the *Goldscore* or the *Chemscore* function; dockings were run on a Linux cluster; (iv) compounds are rescored with *Goldscore*, *Chemscore*, and *Drugscore*. Each virtual screen was repeated five times to minimize the effect of the stochastic nature of the docking algorithm. Merged libraries were generated for all 25 combinations of the five virtual screening runs of the *active compound library* and the five runs for the corresponding *focused compound library*; each merged library was ranked and the 25 ranked lists were averaged. Average, single-processor CPU times were 1.7 min per compound for *Goldscore* dockings and 0.7 min per compound for *Chemscore* dockings on 1GHz/PentiumIII PC's running Linux.

3. RESULTS AND DISCUSSION

Single-Scoring-Function Ranking. Figure 4 shows the virtual screening results, using-single-scoring-function ranking, for the 6 validation sets. It is clear that significant enrichments are obtained for all targets. If we focus on the top 1% of the databases, the maximum enrichments vary between 10-fold for the low-molecular-weight cdk2 binders and 60-fold for the estrogen agonists and antagonists. It is interesting to note that the best results are obtained for the test sets for which the known binders have the highest affinities, the estrogen receptor agonists and antagonists, and the lowest enrichments are obtained for the test set containing the weakest binders, the low-molecular-weight cdk2 set.

The performance of the scoring protocols used is very much target dependent. The *Goldscore* function appears to perform best when ligand binding is mainly hydrogen-bond driven. In neuraminidase, the negatively charged parts of the ligands are locked into place by three arginine residues, and other parts of the ligand form additional hydrogen bonds with the target. For this target *Goldscore* clearly outperforms the other two scoring functions. *Goldscore* also performs well against ptp1b, where ligands all form hydrogen bonds with the "cradle" of N−H groups at the bottom of the phosphate-binding pocket. It is interesting to note that *Goldscore* performs significantly better for estrogen receptor antagonists than for the agonists. Although the estrogen receptor binding site is mainly hydrophobic, the agonists typically form two hydrogen bonds at either side of the pocket. Antagonists, on the other hand often form an additional salt bridge near the entrance of the binding site. This extra hydrogen-bonding interaction could be the reason *Goldscore* works better for antagonists. Stahl and Rarey noticed that the FlexX scoring function (which also performs best when protein−ligand binding is hydrogen bond driven) ranks antagonists highly, compared to agonists. This difference could also be caused by the additional salt bridge formed by many antagonists, although the fact that antagonists are significantly bigger than agonists may also have played a role (see above). It is interesting to note here that the *Goldscore* function rewards ionic hydrogen bonds much more than neutral ones, whereas the *Chemscore* function rewards all hydrogen bonds the same. We ensured that the numbers of donors and acceptors of the compounds in the focused libraries were similar to those of the known binders, but we did not specifically ensure that the molecular charges were similar. For targets with a strong ionic component to the interaction between protein and ligand, this could have helped the enrichments obtained with the *Goldscore* function (and possibly the *Drugscore* function), but not those obtained with the *Chemscore* function.

Chemscore performs best for cdk2 and the estrogen receptor, both targets for which a significant part of the affinity of the ligands can be attributed to lipophilic interactions. The enrichments obtained for the estrogen receptor, with the *Chemscore* function, are the highest we observed in any of our test sets: 60-fold in the top 1% of the databases. These results are similar to those reported by Waszkowycz et al.;[51] they obtained a 76-fold enrichment in the top 1%, when they docked 1.1 million compounds against the estrogen receptor, using exactly the same set of 20 estrogen receptor agonists, and using the *Chemscore* function. For cdk2, the *Chemscore* function is the only function that gives reasonable enrichments. To a significant extent, this can be attributed to the C−H···A term we added to the *Chemscore* function. The reason this term has an effect on the cdk2 enrichments is because many of the ligands in our test sets form C−H···A interactions with cdk2. Figure 1 shows an example of a cdk2 complex that contains C−H···A interactions. Invariably, these interactions involve the backbone carbonyl oxygens of Glu81 and Leu83.

Overall, the *Drugscore* function performs the worst of the three scoring functions tested here. It gives significant enrichments for both the estrogen receptor agonists and antagonists, but for the other targets the results obtained with the *Drugscore* function are not much better than what is expected at random. These results are in line with those obtained by Stahl and Rarey; they concluded that *Drugscore* can model lipophilic interactions well but struggles when protein−ligand binding is driven mainly by hydrogen bonding. Interestingly, where the *Goldscore* function performs better than *Drugscore* for estrogen receptor antagonists, *Drugscore* works better than *Goldscore* for the agonists. It appears that *Drugscore* struggles to discriminate between compounds that do and those that do not form the additional salt bridge formed by many of the antagonists.

Consensus Ranking. Our approach to assess the effectiveness of consensus ranking is analogous to that used by Stahl and Rarey,[15] i.e., the enrichments are compared with those obtained using the individual scoring functions. Table 3 compares the performance of the three consensus-ranking

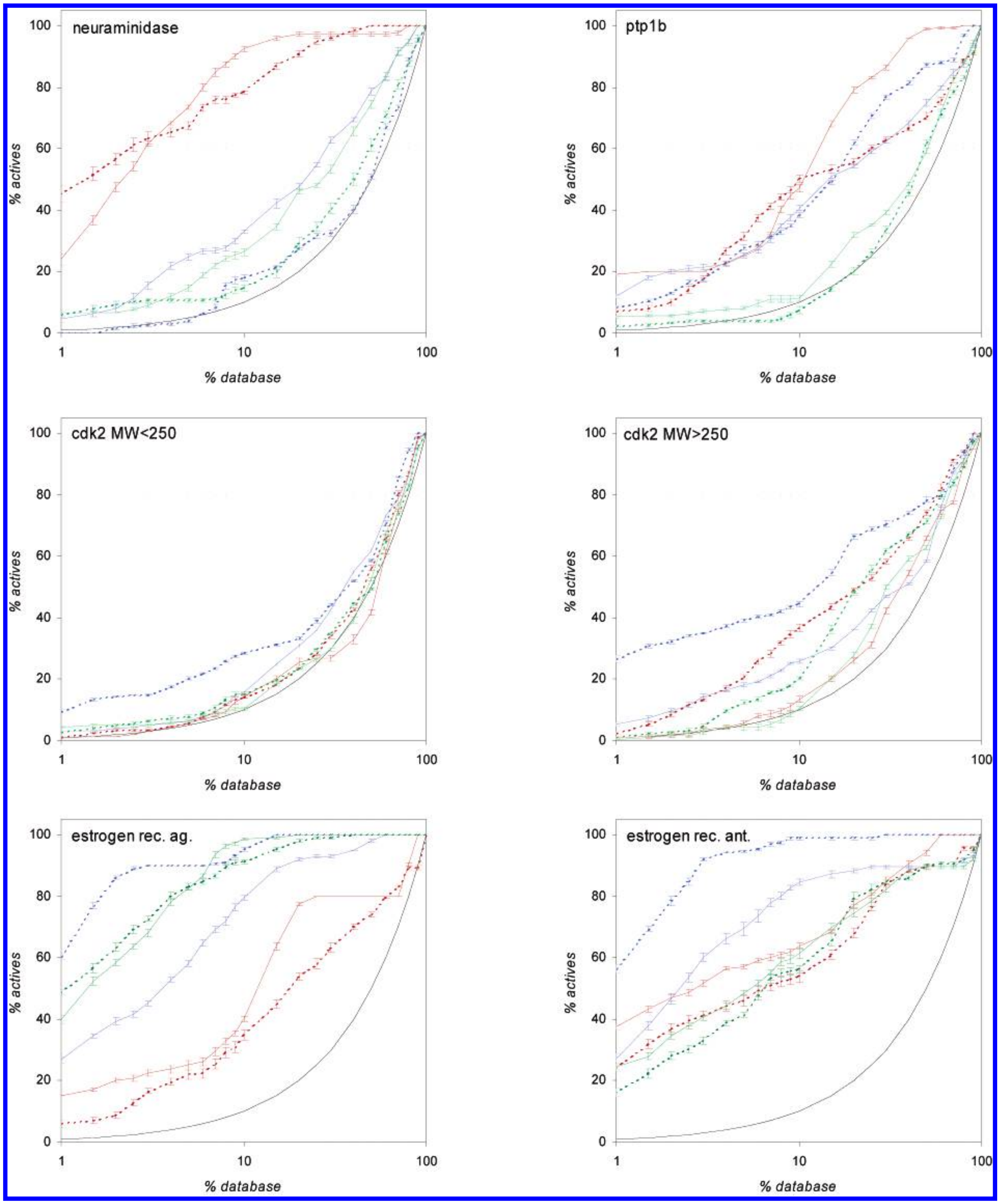

Figure 4. Enrichment graphs for the six validation sets studied. All results are the averages over the 25 combinations of the five runs for the actives and the five runs for the *focused library*. The error bars represent the errors in the mean. Runs for which the *Goldscore* function was used to drive the dockings are shown as solid lines; those for which the dockings were generated with *Chemscore* are shown as dotted lines. Results are shown for rescoring and ranking with *Goldscore* (red), *Chemscore* (blue), and *Drugscore* (green). The black line represents the fraction of actives expected at random.

strategies tested. It appears that, on average, consensus ranking does not perform as well as the best of the individual scoring-functions-combined. However, the performance of consensus ranking *is* generally better than the average performance of the individual scoring-functions-combined. Interestingly, in our validations, the performance of the three consensus ranking strategies is *rank-by-number* > *rank-by-rank* > *rank-by-vote*, which is exactly what was predicted by the simulated virtual screening experiments done by Wang and Wang.[47]

Figure 5 summarizes the performance of the best-performing consensus ranking protocol, *rank-by-number*, for all 6 validation sets and all single-scoring-function and consensus ranking protocols tested. As was also observed by Stahl and Rarey,[15] Figure 5 shows that in some cases consensus ranking (slightly) outperforms the individual scoring functions;

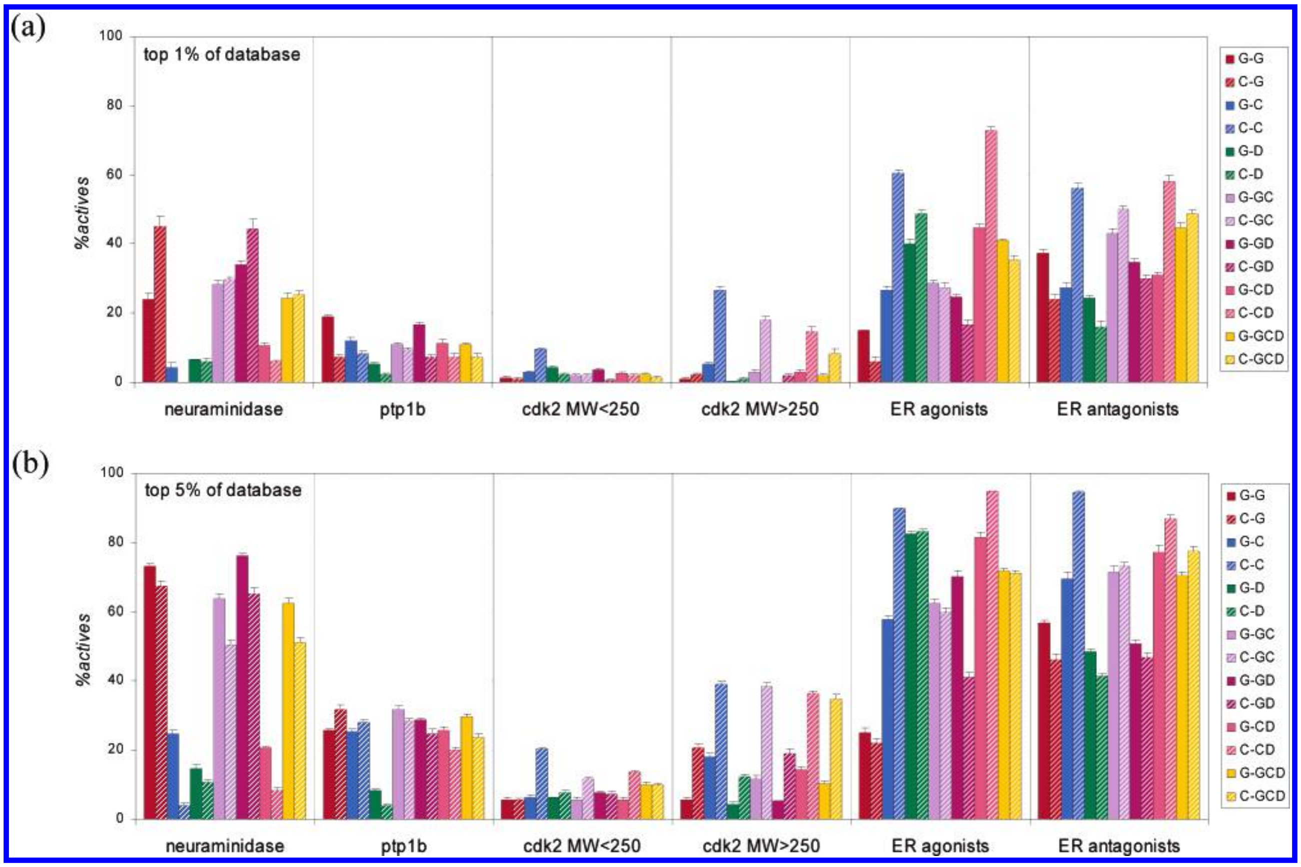

Figure 5. Overview of consensus ranking using the *rank-by-number* protocol. Results are shown for (a) the top 1% of the database and (b) the top 5% of the database. The first character in the figure legend indicates the scoring function used for docking ("G" for *Goldscore*, "C" for *Chemscore*, and "D" for *Drugscore*), and the second character is the scoring function used for scoring and ranking.

Table 3. Comparison of the Performance of Three Consensus-Ranking Protocols[a]

	top 1% of database		top 5% of database	
	$\langle f_{\mathrm{cons}}/f_{\mathrm{max}}\rangle$[b]	$\langle f_{\mathrm{cons}}/f_{\mathrm{av}}\rangle$[b]	$\langle f_{\mathrm{cons}}/f_{\mathrm{max}}\rangle$[b]	$\langle f_{\mathrm{cons}}/f_{\mathrm{av}}\rangle$[b]
rank-by-number	0.78(5)	1.18(7)	0.89(3)	1.24(4)
rank-by-rank	0.75(7)	1.09(9)	0.80(5)	1.07(4)
rank-by-vote	0.69(7)	1.00(9)	0.75(5)	1.00(5)

[a] All values are averages over the six validation sets and the eight consensus ranking protocols, giving a total number of 48 observations for each value presented; standard deviations are given in parentheses. [b] f_{cons} is the fraction of actives retrieved using consensus ranking; f_{max} is the fraction of actives retrieved by the best of the individual scoring-functions-combined; f_{av} is the fraction of actives retrieved, averaged over the individual scoring-functions-combined.

typically, however an individual single scoring function gives the best performance. Consensus ranking is mainly useful when there is limited knowledge about the target and its inhibitors. In such cases, consensus ranking provides a safer and more robust strategy to rank compounds than ranking based on a single scoring function (see also ref 12).

The fact that, typically, consensus ranking does not perform better than the best of the individual scoring-functions-combined is an important observation. It was noticed by Stahl and Rarey,[15] and, on close inspection, the consensus ranking data presented by Charifson et al.[12] also shows that this is the case. But the simulated virtual screening experiments by Wang and Wang seem to show that, in theory, combining multiple scoring functions should always give improved performance over individual scoring functions. We believe the main reason the simulation results deviate from what is observed in practice is because in their models Wang and Wang assume that the accuracy of each scoring function combined is the same (in their model, the standard error in the predicted affinity, s, is 2.0 log units, for each scoring function). In Figure 6 we have attempted to reproduce Wang and Wang's simulation, using the *rank-by-number* consensus ranking strategy (the details of the methods used are given by Wang and Wang[47]). The dotted red line represents our repeat of the simulation done by Wang and Wang, i.e., where $s = 2.0$ for all scoring-functions-combined. It is clear that, in such a case, adding extra scoring functions to the consensus ranking protocol always improves performance. But typically the performance varies for different scoring functions. The solid blue line in Figure 6 shows what happens when the first scoring function performs very well ($s = 1.0$), but the performance of the other scoring functions is poorer ($s = 3.0$). In this case, adding the second scoring function mainly adds noise, and the performance drops, relative to that of the first scoring function. We believe this explains why, typically, consensus ranking does not work better than the best of the individual scoring-functions-combined. It also explains why, when the individual scoring-functions-combined perform similarly, consensus ranking *can* improve performance. In this context, it may be worth considering assigning different weights to different scoring functions, depending on the standard error in the predicted affinities.

Pharmacophore Restraints. It is clear from Table 4 and Figure 4 that the most difficult of the virtual screening test sets presented here are the cdk2 sets, particularly the low-molecular-weight set. Structural knowledge about the way

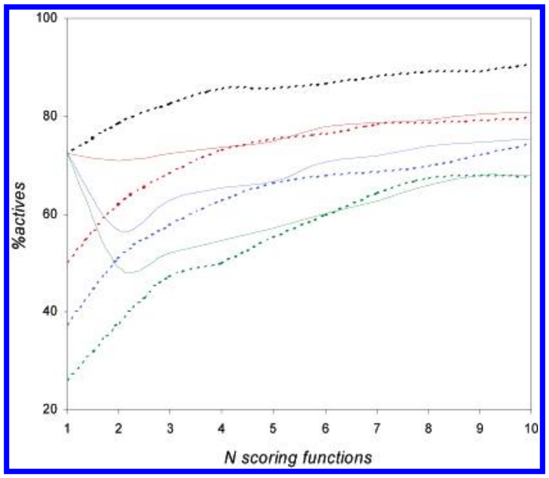

Figure 6. Simulated effect of consensus ranking with an increasing number of scoring functions, using the rank-by-number protocol, following Wang and Wang.[47] The dotted lines represent cases where the standard deviation in the predicted affinity is the same for all scoring-functions-combined; results are shown for $s = 1.0$ (black), $s = 2.0$ (red), $s = 3.0$ (blue), and $s = 4.0$ (green); in these cases, consensus ranking is seen to be helpful. The solid lines represent simulations where $s = 1.0$ for the first scoring function, but for the other scoring functions, s is greater than 1.0; results are shown where s for scoring functions 2–10 is 2.0 (red), 3.0 (blue), and 4.0 (green); it can be seen how in such cases consensus ranking can reduce the enrichment rates, compared to the most accurate single scoring function.

Table 4. Success Rates[a] for Binding Mode Predictions against the Six Validation Sets, Using *Goldscore* and *Chemscore*[b]

	N	*Goldscore*	*Chemscore*
neuraminidase	15	96(4)/96(4)[c]	63(6)/89(7)[c]
ptp1b	5	76(9)	56(30)
cdk2 MW<250	18	22(7)/59(4)[c]	48(6)/65(6)[c]
cdk2 MW>250	17	33(3)	65(4)
ER agonists	3	53(30)	80(18)
ER antagonists	2	90(22)	60(22)

[a] Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode. [b] All values are averages over the five runs used in the virtual screening validation experiments (Figure 4); standard deviations are given in parentheses. [c] Docked under pharmacophore restraints.

compounds bind to cdk2 could possibly be of help in this case. Cdk2 inhibitors invariably form a hydrogen bond with the backbone N–H of Leu83. Hence, we constructed a pharmacophore consisting of a single *acceptor* point at the exact position of the ring nitrogen atom in AT934 (see Figure 1). We used a block-shaped overlap function with a radius of 0.5 Å for the acceptor point. This pharmacophore was used as a restraint to the *Chemscore* function during the docking of all compounds, with a pharmacophore weight of $C_p = 30$. At the ranking stage, however, the pharmacophore term was not taken into account.

It is clear from Table 4 that the pharmacophore significantly improves the binding mode predictions for the known binders. The improvement in the enrichments is also significant, although not huge (see Figure 7). The remaining problem with the test set for the low-molecular-weight cdk2 actives is that it is often hard, even for a trained molecular modeler, to distinguish actives from the high-ranking library compounds. This could well be related to the fact that hit rates for cdk2 are typically relatively high; i.e., some of the high-scoring library compounds could actually prove to be actives. The fact that, for such a challenging test set like this, we can get up to 14-fold enrichment in the top 1% of the database is very encouraging.

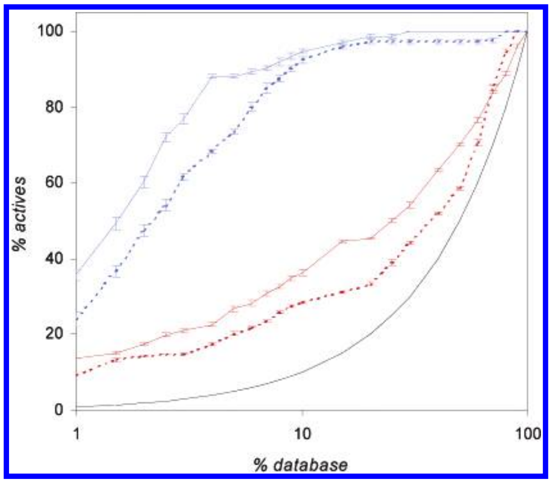

Figure 7. Effect of pharmacophore restraints on enrichment rates. The dotted lines represent validation runs without pharmacophore restraints; the solid lines represent runs for which pharmacophore restraints were used to guide the dockings. Results are shown for the low-molecular-weight cdk2 compounds (red) and for the neuraminidase compounds (blue). The black line represents the fraction of actives expected at random.

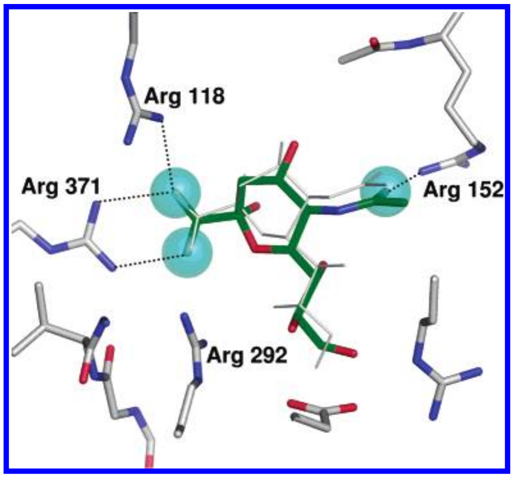

Figure 8. X-ray structure of sialic acid, in complex with neuraminidase (PDB entry 1mwe). The three pharmacophore acceptor points used to guide the dockings are shown in cyan. In gray, a typical *Goldscore* docking for sialic acid is shown.

The other target for which the actives form a conserved motif of interactions with the binding site is neuraminidase. All neuraminidase actives have at least two acceptors in the pocket surrounded by Arg118, Arg292, and Arg371 and a third acceptor interacting with Arg152. We constructed a pharmacophore consisting of three block-shaped acceptor points (see Figure 8), each with a radius of 1.0 Å (the variance in the acceptor positions is slightly bigger than in the case of cdk2). This pharmacophore was then used as a restraint to the *Goldscore* function, with a pharmacophore weight of $C_p = 30$, to dock all compounds; again, the pharmacophore term was not taken into account for ranking the compounds.

In this case, the pharmacophore restraint does not help the prediction of the binding modes (see Table 4); *Goldscore* predicts nearly all of the binding modes of the known actives correctly, without the restraint. However, the enrichment is

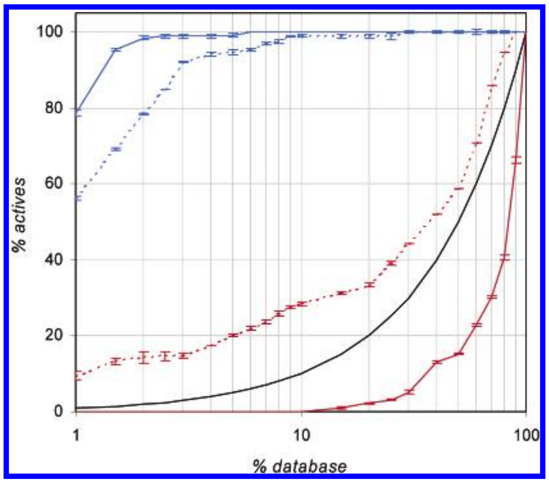

Figure 9. Effect of the type of library on enrichment rates. The dotted lines represent validation runs against *focused libraries*; the solid lines represent runs against random, drug-like libraries. Results are shown for the low-molecular-weight cdk2 compounds (red) and for the estrogen receptor antagonists (blue). The black line represents the fraction of actives expected at random.

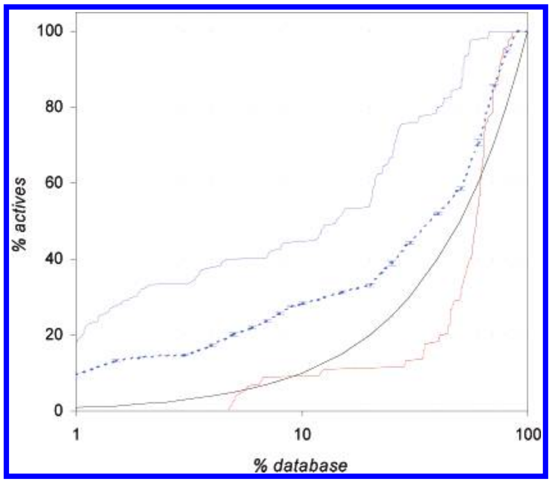

Figure 10. Effect of the quality of the binding modes on enrichment rates for the low-molecular-weight cdk2 compounds. The solid blue line represents actives for which the binding mode is predicted correctly (i.e. within 2.0 Å RMSD of the experimental binding mode); the solid red line represents actives for which the binding mode is not predicted correctly. The dotted blue line represents all actives (see Figure 4). The black line represents the fraction of actives expected at random.

significantly improved as a result of the pharmacophore (see Figure 7). This effect must be caused by the fact that the pharmacophore restraint forces the library compounds to adopt less favorable binding modes, hence shifting them to lower ranks.

Library Generation. Throughout this work, we have tested the scores and ranks of the known binders against *focused libraries*, i.e., libraries that contain compounds with 1D properties similar to those of the known actives. In the Methodology section we used simulated virtual screening experiments to illustrate why it is important to use such *focused libraries*. Figure 9 shows how enrichments can be affected in actual virtual screens, when inappropriate libraries are used. The dotted lines represent virtual screens done against *focused libraries*; the solid lines represent screens against *random libraries* (these libraries contain 100 compounds for each known binder, selected randomly from the compounds in ATLAS that pass Lipinski's Rule of Five[38,39]). The results shown in Figure 9 are for the validation set with the largest known binders (the estrogen receptor antagonists) and for the validation set with the smallest known binders (the low-molecular-weight cdk2 binders). In both cases all compounds were docked and ranked with the *Chemscore* function.

On average, the estrogen receptor antagonists are larger than the ATLAS compounds; furthermore, the other 1D properties of the ATLAS compounds will typically not be similar to those of known estrogen receptor antagonists. As a result, the enrichments obtained against the *random library* are significantly higher than those against the *focused library*. The low-molecular-weight cdk2 binders, on the other hand, are much smaller than the average ATLAS compound. The unspecific interactions formed by the larger compounds in the *random library* outweigh the smaller number of specific interactions formed by the low-molecular-weight cdk2 binders, as would be expected from Figure 3. The result is that *inverse* enrichments (i.e. below 1.0) are obtained for this set of actives, when screened against the *random library*. It is interesting to note the qualitative similarity between Figure 3c and Figure 9.

An alternative to using *focused libraries* in order to eliminate molecular size bias is to scale the scores, according to the size of the compounds. Pan et al.[25] suggested to divide the score of each compound by N^x, where N is the number of heavy atoms in the compound and x is a constant (the authors tested $x = 1/3$, $x = 1/2$, $x = 2/3$, and $x = 1$); the value of x determines the size of the compounds retrieved. The difficulty with using this approach for virtual screening validation is that the molecular size distribution of the actives (i.e. the desired size of the retrieved compounds) varies between test sets (see Figure 2). Therefore, in our opinion, the value of x would need to be reoptimized for each set of known binders. Moreover, the approach does not take into account biases in 1D properties, other than molecular size. The same is true for the approach taken by Ewing et al., who suggested to add a fixed penalty per atom to the score in order to eliminate the molecular size bias.[52]

Accuracy of Binding Modes. Table 4 shows the performance of the two docking protocols used (*Goldscore* and *Chemscore*), in terms of their ability to reproduce the binding modes of the actives for which X-ray structures are available. It has to be pointed out that, because the numbers of X-ray structures that are available are low, the uncertainties in the success rates presented in Table 4 are high. Still, they are roughly in line with the success rates obtained against the CCDC/Astex validation set. What is also quite clear is that the performance of both scoring functions are target dependent. *Goldscore* appears to perform better on neuraminidase (something we also observed in previous work[36]), and the *Chemscore* function performs better for cdk2.

For the low-molecular-weight cdk2 binders, we have a sufficient number of X-ray structures to test if there is a correlation between the enrichments and the quality of the binding modes. Figure 10 again shows the enrichment graph for this test set when docking and scoring is done with *Chemscore* (dotted blue line). In addition, it shows the enrichment graphs for the subset of actives for which the RMSD with respect to the X-ray structure is less than 2.0 Å (blue line) and for the subset for which the RMSD is above

2.0 Å (red line). Enrichments are significantly higher for the actives for which the binding modes are predicted correctly; no enrichment is obtained for the actives that have been docked incorrectly. Similar results are obtained for the high-molecular-weight cdk2 binders and the neuraminidase actives (not shown).

These results clearly show that, to get good enrichments against focused libraries such as the ones we use here, it is important that the docking program used produces high-quality binding modes. There have been examples in the literature where authors have found no correlation between the quality of the dockings and the enrichments.[11] This may indicate that the obtained enrichments are little to do with the 3D structure of the target but may be related to a difference in a lower dimension between the actives and the library compounds. The actives could, for example, score better because they are larger than the average compound in the library (see above). Or, if the binding site is mainly lipophilic, the actives are likely to be more hydrophobic than the library compounds and would score better against the target, whether the binding modes produced are correct or not. In such cases, good enrichments could be obtained by simple 1D screening, e.g. on the molecular weight or on the ClogP of the compounds.

Some of the results presented in Figure 4 also illustrate that good binding modes improve enrichments. For example, we know from Table 4 that *Chemscore* produces much better binding modes than *Goldscore* for cdk2 binders. In this light, it is interesting to note that ranking with the *Chemscore* function only performs well for cdk2 when the dockings are produced with *Chemscore*; there is virtually no enrichment when the dockings produced with *Goldscore* are scored and ranked with *Chemscore*. For the high-molecular-weight cdk2 binders, something similar appears to happen when the *Goldscore* function is used for ranking; in that case the performance is also better when the dockings are produced with *Chemscore*.

It is worth pointing out here that, although we (like most authors) have used an RMSD cutoff of 2.0 Å to separate "correct" binding modes from "incorrect" ones, this can sometimes be misleading.[53] For example, if the key part of a molecule is docked correctly, but its lipophilic tail is docked in subpocket A, rather than in subpocket B, its RMSD may well be above 2.0 Å. However, we do not believe that the use of a different measure for the "correctness" of the predicted binding modes would have affected the conclusions of our analyses.

4. CONCLUSIONS

We have shown that virtual screening with GOLD can give significant enrichments (10−60-fold) against all six validation sets presented here. As was observed by other workers, performance depends on the scoring function(s) used and on the target. To enhance performance against cdk2, a term that takes into account certain types of C−H···A interactions was added to the *Chemscore* function.

A systematic analysis of three consensus-ranking protocols showed that the *rank-by-number* approach is most effective; this provides experimental support for the results obtained from simulated virtual screens done by Wang and Wang.[47] Typically, consensus ranking does not perform better than the best of the individual scoring-functions-combined, but it does provide a more robust ranking method when the performance of the individual scoring functions is unknown (see also refs 12, 15, and 26).

We have shown that actives need to be screened against *focused libraries*, i.e., containing compounds with 1D properties similar to those of the actives; *random libraries*, or even drug-like compound libraries can cause meaningless enrichments, which, in certain cases, could also have been obtained by much simpler ligand-based screening techniques. Also, clear evidence was presented that good-quality binding modes are a prerequisite for good enrichments against focused libraries. Pharmacophore restraints were incorporated into GOLD and were shown to give increased enrichments against cdk2 and neuraminidase.

Finally, we have shown that protein−ligand docking is not only useful for the identification of nanomolar binders but can also be effective for the screening for weakly-binding fragment-like hits. For the low-molecular-weight cdk2 set (actives with molecular weights below 250 and weak-to-moderate affinities) up to 14-fold enrichments can be obtained in the top 1% of the database. Similarly, for the ptp1b validation set (which also contains a significant number of weak, fragment-like binders), enrichments of 15−20-fold can be obtained.

REFERENCES AND NOTES

(1) Makino, S.; Kuntz, I. D. Automated flexible ligand docking method and its application for database search. *J. Comput. Chem.* **1997,** *18*, 1812−1825.

(2) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996,** *261*, 470−489.

(3) 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.

(4) 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*, 727−748.

(5) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002,** *16*, 151−166.

(6) Tondi, D.; Slomczynska, U.; Costi, M. P.; Watterson, D. M.; Ghelli, S.; Shoichet, B. K. Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase. *Chem. Biol.* **1999,** *6*, 319−331.

(7) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002,** *45*, 2213−2221.

(8) Brenk, R.; Naerum, L.; Gradler, U.; Gerber, H.-D.; Garcia, G. A.; Reuter, K.; Stubbs, M. T.; Klebe, G. Virtual Screening for Submicromolar Leads of tRNA-guanine Transglycosylase Based on a New Unexpected Binding Mode Detected by Crystal Structure Analysis. *J. Med. Chem.* **2003**, *46*, 1133−1143.

(9) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7*, 1047−1055.

(10) Good, A. Structure-based virtual screening protocols. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 301−307.

(11) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(12) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(13) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol Graph. Model.* **2002**, *20*, 281−295.

(14) Baxter, C. A.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. New approach to

molecular docking and its application to virtual screening of chemical databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 254−262.

(15) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(16) Jenkins, J. L.; Kao, R. Y.; Shapiro, R. Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins* **2003**, *50*, 81−93.

(17) Good, A. C.; Cheney, D. L.; Sitkoff, D. F.; Tokarski, J. S.; Stouch, T. R.; Bassolino, D. A.; Krystek, S. R.; Li, Y.; Mason, J. S.; Perkins, T. D. Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. *J. Mol. Graph. Model.* **2003**, *22*, 31−40.

(18) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.

(19) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 899−907.

(20) Verdonk, M. L.; Cole, J. C.; Hartshorn, M.; Murray, C.; Taylor, R. D. Improved protein−ligand docking using GOLD. *Proteins* **2003**, *52*, 609−623.

(21) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. 1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(22) Pierce, A. C.; Sandretto, K. L.; Bemis, G. W. Kinase inhibitors and the case for CH···O hydrogen bonds in protein−ligand binding. *Proteins* **2002**, *49*, 567−576.

(23) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(24) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discov. Des.* **2000**, *20*, 115−144.

(25) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.

(26) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(27) Fradera, X.; Knegtel, R. M. A.; Mestres, J. Similarity-driven flexible ligand docking. *Proteins* **2000**, *40*, 623−636.

(28) Hindle, S. A.; Rarey, M.; Buning, C.; Lengaue, T. Flexible docking under pharmacophore type constraints. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129−149.

(29) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002**, *1*, 45−54.

(30) Boehm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **2000**, *43*, 2664−2674.

(31) Carr, R.; Jhoti, H. Structure-based screening of low-affinity compounds. *Drug Discov. Today* **2002**, *7*, 522−527.

(32) Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* **1999**, *6*, 755−769.

(33) Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. NMR-based screening in drug discovery. *Q. Rev. Biophys.* **1999**, *32*, 211−240.

(34) Nienaber, V. L.; Richardson, P. L.; Klighofer, V.; Bouska, J. J.; Giranda, V. L.; Greer, J. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* **2000**, *18*, 1105−1108.

(35) O'Reilly, M.; Woolford, A. J.-A. 2003, Unpublished work.

(36) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 855−869.

(37) Iversen, L. F.; Andersen, H. S.; Branner, S.; Mortensen, S. B.; Peters, G. H.; Norris, K.; Olsen, O. H.; Jeppesen, C. B.; Lundt, B. F.; Ripka, W.; Moller, K. B.; Moller, N. P. Structure-based design of a low molecular weight, nonphosphorus, nonpeptide, and highly selective inhibitor of protein-tyrosine phosphatase 1B. *J. Biol. Chem.* **2000**, *275*, 10300−10307.

(38) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(39) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3−26.

(40) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367−382.

(41) Hartshorn, M. J.; Watson, P. 2002, Unpublished work.

(42) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic assignment of bond and atom types for protein ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.

(43) Sayle, R. PDB: Cruft to content (perception of molecular connectivity from 3D coordinates) (http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html) 2001.

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

(45) Sotriffer, C. A.; Gohlke, H.; Klebe, G. Docking into knowledge-based potential fields: a comparative evaluation of DrugScore. *J. Med. Chem.* **2002**, *45*, 1967−1970.

(46) Daylight Chemical Information Systems, Mission Vieho, CA. (http://www/daylight.com). 2003.

(47) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(48) Watson, P.; Verdonk, M. L.; Hartshorn, M. A web-based platform for virtual screening. *J. Mol. Graph. Model.* **2003**, *22*, 71−82.

(49) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1998**, *28*, 31−36.

(50) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Method.* **1990**, *3*, 537−547.

(51) Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM. Syst. J.* **2001**, *40*, 360−376.

(52) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(53) Smith, R.; Hubbard, R. E.; Gschwend, D. A.; Leach, A. R.; Good, A. C. Analysis and optimization of structure-based virtual screening protocols. (3). New methods and old problems in scoring function design. *J. Mol. Graph. Model.* **2003**, *22*, 41−53.

CI034289Q