Improved Protein–Ligand Docking Using GOLD

Marcel L. Verdonk,[1*] Jason C. Cole,[2] Michael J. Hartshorn,[1] Christopher W. Murray,[1] and Richard D. Taylor[1]
[1]*Astex Technology, Ltd., Cambridge, United Kingdom*
[2]*Cambridge Crystallographic Data Centre, Cambridge, United Kingdom*

ABSTRACT **The Chemscore function was implemented as a scoring function for the protein–ligand docking program GOLD, and its performance compared to the original Goldscore function and two consensus docking protocols, "Goldscore-CS" and "Chemscore-GS," in terms of docking accuracy, prediction of binding affinities, and speed. In the "Goldscore-CS" protocol, dockings produced with the Goldscore function are scored and ranked with the Chemscore function; in the "Chemscore-GS" protocol, dockings produced with the Chemscore function are scored and ranked with the Goldscore function. Comparisons were made for a "clean" set of 224 protein–ligand complexes, and for two subsets of this set, one for which the ligands are "drug-like," the other for which they are "fragment-like." For "drug-like" and "fragment-like" ligands, the docking accuracies obtained with Chemscore and Goldscore functions are similar. For larger ligands, Goldscore gives superior results. Docking with the Chemscore function is up to three times faster than docking with the Goldscore function. Both combined docking protocols give significant improvements in docking accuracy over the use of the Goldscore or Chemscore function alone. "Goldscore-CS" gives success rates of up to 81% (top-ranked GOLD solution within 2.0 Å of the experimental binding mode) for the "clean list," but at the cost of long search times. For most virtual screening applications, "Chemscore-GS" seems optimal; search settings that give docking speeds of around 0.25–1.3 min/compound have success rates of about 78% for "drug-like" compounds and 85% for "fragment-like" compounds. In terms of producing binding energy estimates, the Goldscore function appears to perform better than the Chemscore function and the two consensus protocols, particularly for faster search settings. Even at docking speeds of around 1–2 min/compound, the Goldscore function predicts binding energies with a standard deviation of ~10.5 kJ/mol. Proteins 2003;52:609–623.**

Key words: scoring functions; consensus docking; virtual screening; chemscore; binding mode; binding affinity

INTRODUCTION

Predicting the binding modes and affinities of compounds when they interact with a protein-binding site lies at the heart of structure-based drug design. Consequently, the number of algorithms available for protein–ligand docking is large. DOCK,[1] FlexX,[2] PRO_LEADS,[3] and GOLD[4,5] are examples of docking programs, but many more are reported in the literature (for an overview of docking strategies see Taylor et al.[6]). Most approaches consider the protein to be (mostly) rigid and allow the ligand to be flexible.

The key characteristic of a good docking program is its ability to reproduce the experimental binding modes of ligands. To test this, a ligand is taken out of the X-ray structure of its protein–ligand complex and docked back into its binding site. The docked binding mode is then compared with the experimental binding mode, and a root-mean-square distance (RMSD) between the two is calculated; a prediction of a binding mode is considered successful if the RMSD is below a certain value (usually 2.0 Å). Recently, Nissink et al. pointed out that to establish the success rate of a docking program, a large and carefully constructed set of protein–ligand complexes is required.[7] As far as we know, the only flexible docking programs that were tested on large test sets are PRO_LEADS, FlexX, and GOLD. PRO_LEADS was demonstrated to give success rates of up to 84% on a test set of 70 complexes[8]; FlexX gave a success rate of 47% on a test set of 200 complexes[9]; and recently, GOLD was shown to give a 68% success rate on the CCDC/Astex validation set of 305 complexes[7]; all complexes in all three test sets were taken from the Protein Data Bank (PDB).[10]

An important use of protein–ligand docking programs is virtual screening, in which large libraries of compounds are docked into a target binding site and scored. For this purpose, the dockings need to be quick. Speeding up a docking protocol is often done at the cost of sampling fewer binding modes, and, as a result, reduces the success rates. It is therefore important that search parameters are chosen that give docking speeds useful for virtual screening applications (in our case, up to 2 min/compound on a single processor), with an acceptable loss in docking accuracy.

Another characteristic of a good docking program is the ability of its scoring function to score and rank ligands according to their experimental binding affinities. To test

*Correspondence to: Marcel L. Verdonk, Astex Technology Ltd., 436 Cambridge Science Park, Milton Road, Cambridge CB4 OQA, UK. E-mail: m.verdonk@astex-technology.com

Received 26 November 2002; Accepted 28 January 2003

this, the predicted binding affinities, or scores, are plotted against the experimental binding affinities; the key indicator for the quality of the predicted affinities is the standard deviation s (or the cross-validated error, s_{press}). For calculations based on the experimental binding modes of the ligands, good scoring functions, suitable for docking, give s values of around 8 kJ/mol,[11,12] which corresponds to about 1.5 orders of magnitude in the affinity. But, if the binding modes are produced by a docking program, the agreement between calculated and predicted binding affinities tends to drop quickly, particularly for faster search protocols (see Baxter et al.[8]). 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 this article, we describe the implementation of the Chemscore function as a scoring function for GOLD and its usefulness to improve docking accuracy and the prediction of binding energy. Chemscore is the scoring function used in PRO_LEADS, and, because higher success rates are quoted for this program than for GOLD (see above), it seemed an obvious improvement to make to GOLD. Moreover, the Chemscore function was parameterized against the experimental binding affinities for a test set of 82 protein–ligand complexes. Notionally, therefore, the Chemscore function should give a better correlation with affinity than the GOLD scoring function (henceforth known as "Goldscore function"), which was not parameterized against binding affinity data.

We first describe the validation of our implementation of the Chemscore function and its partial reoptimization for better performance with "raw" PDB files. Then we analyze the performance of the Chemscore function against the original GOLD scoring function, in terms of docking accuracy, ability to predict binding affinities, and speed of the dockings. In parallel, we investigate the performance of two combined, or "consensus," docking protocols that make use of both the Goldscore and the Chemscore function; in the first protocol, "Goldscore-CS," the dockings produced with the Goldscore function are rescored and *reranked* with the optimized Chemscore function, and in the second protocol, "Chemscore-GS," the dockings produced with the optimized Chemscore function are rescored and *reranked* with the Goldscore function. All results are presented for a "clean" subset of 224 complexes from the CCDC/Astex validation set (see Materials and Methods section), and for two subsets of this "clean list": one in which the ligands in the complexes are "drug-like," and the other in which they are "fragment-like" (see Materials and Methods section for definitions). In analogy with Verkhivker et al.,[13] we classify the complexes for which GOLD fails to predict the experimental binding mode successfully, as "soft" or "hard" failures (see Materials and Methods section). Docking accuracy and quality of the predicted binding affinities are analyzed for various search settings (i.e., varying docking speeds), and recommendations are made.

MATERIALS AND METHODS

The CCDC/Astex Validation Set

For all the studies presented here, we use the CCDC/Astex validation set, published recently by Nissink et al.[7] This set of 305 protein–ligand complexes from the PDB is currently the largest test set for docking programs, and great care was taken to ensure that the protonation states and bond types of ligand and protein are correct. The "clean list," a subset of the validation set, consists of 224 complexes; these complexes do not exhibit protein–ligand clashes, crystallographic contacts, or unlikely ligand geometries; for closely related complexes, only one representative was kept, making the "clean list" also more diverse than the overall validation set. In this study, we focus on the "clean list."

The CCDC/Astex validation set contains a wide range of ligands, varying from very small to quite large peptidic compounds. This is a good basis for the validation of a docking program, but not all of these ligands are representative of structure-based drug design compounds. Therefore, we also test the performance of GOLD on a subset of 139 complexes, for which the ligands are more "drug-like." Various measures of "drug-likeness" can be found in the literature. Lipinski's original analysis of orally administered drugs and drug candidates led to the well-known Rule of Five.[14,15] Here, we use the simpler rules to predict bioavailability derived by Veber et al.[16] Hence, in our "drug-like list," only complexes from the "clean list" were included where the ligand has 10 or fewer rotatable bonds and a polar surface area $\leq$140 Å^2. We would like to stress that because we did not use any additional filters (e.g., to remove complexes with toxic or reactive ligands), "drug-likeness" is here limited to bioavailability. Recently, screening small, "fragment-like" compounds has become a popular strategy for hit-identification,[17–22] particularly because smaller leads are generally easier to optimize. Hence, we also list the performance of GOLD for a subset of 79 complexes for which the ligands are small and "fragment-like." In this "fragment list," only complexes from the "clean list" were included where the ligand is not covalently bound to the protein, has five or fewer rotatable bonds, and between 7 and 20 nonhydrogen atoms.

GOLD

Most parts of the GOLD program have been described by Jones et al.[4,5] Like all other docking programs, GOLD consists of three main parts:

1. A *scoring function* to rank different binding modes; the Goldscore function is a molecular mechanics–like function with four terms:

$$\text{GOLD Fitness} = S_{hb\_ext} + S_{vdw\_ext} + S_{hb\_int} + S_{vdw\_int}, \quad (1)$$

where S_{hb_ext} is the protein–ligand hydrogen-bond score and S_{vdw_ext} is the protein-ligand van der Waals score. S_{hb_int} is the contribution to the *Fitness* due to intramolecular hydrogen bonds in the ligand; this term is switched off in all calculations presented in this work

TABLE I. GA Settings

GA setting	$N_{dockings}$	N_{ops}	Early termination
Exhaustive	100	100,000	No
GOLD default 1[a]	20	100,000	Yes
GOLD default 2	10	50,000	Yes
GOLD default 3	10	30,000	Yes
GOLD default 4	10	10,000	Yes
GOLD library	10	1000	Yes

[a]These settings were used by Nissink et al.[7]

(this is the GOLD default, and generally gives the best results); S_{vdw_int} is the contribution due to intramolecular strain in the ligand.

2. A *mechanism for placing the ligand* in the binding site; GOLD uses a unique method to do this, which is based on fitting points; it adds fitting points to hydrogen-bonding groups on protein and ligand, and maps acceptor points on the ligand on donor points in the protein and vice versa. Additionally, GOLD generates hydrophobic fitting points in the protein cavity onto which ligand CH groups are mapped.
3. A *search algorithm* to explore possible binding modes; GOLD uses a genetic algorithm (GA) in which the following parameters are modified/optimized: (a) dihedrals of ligand rotatable bonds; (b) ligand ring geometries (by flipping ring corners); (c) dihedrals of protein OH groups and NH_3^+ groups; and (d) the mappings of the fitting points (i.e., the position of the ligand in the binding site). Of course, at the start of a docking run, all these variables are randomized.

Genetic Algorithm Settings

The GA settings directly affect the timings of a docking run and the likelihood of finding the global optimum. The main parameters that affect timings and accuracy are the number of dockings and the number of GA operations in each docking. The values for these two parameters for the different GA settings used in this work are listed in Table I. Additional GA parameters are taken from the default GOLD settings, and are available as supplementary material.

A GOLD run can be terminated early if the top-ranked binding mode is produced repeatedly. By default, if this option is switched on, GOLD terminates when the top three dockings are within 1.5 Å of each other. The early termination flags are listed in Table I for the various GA settings used in this work.

Local Scoring

To calculate the scores of ligands bound to a target binding site in the experimental binding mode (or any other mode), two options were added to GOLD: (1) Standard *scoring,* which simply takes a binding mode (including orientations of protein OH and NH_3^+ groups) and scores it; and (2) *local optimization,* which performs a normal GOLD (GA) run, but all atoms are kept fixed except terminal protein and ligand OH and NH_3^+ groups. These are allowed to spin around to optimize hydrogen bonds. As for any GOLD run, both options can be 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.

TABLE II. Chemscore Parameters Used in This Study, Except Where Indicated Otherwise

ΔG_o	−5.480 (kJ/mol)	$r_{m,1}$	2.60 (Å)
ΔG_{hbond}	−3.340 (kJ/mol)	$r_{m,2}$	3.00 (Å)
ΔG_{metal}	−6.030 (kJ/mol)		
ΔG_{lipo}	−0.117 (kJ/mol)	$r_{l,1}$	4.10 (Å)
ΔG_{rot}	2.560 (kJ/mol)	$r_{l,2}$	7.10 (Å)
C_{cov}	0.25	r_{clash} (donor–acceptor)	1.60 (Å)
		r_{clash} (metal–acceptor)	1.38 (Å)
r_o	1.85 (Å)	r_{clash} (sulphur)	3.35 (Å)
Δr_1	0.25 (Å)	r_{clash} (other)	3.10 (Å)
Δr_2	0.65 (Å)		
α_0	180 (°)	$\sigma_{f(\Delta r_{DA},\ \Delta r_1,\ \Delta r_2)}$	0.10 (Å)
$\Delta\alpha_1$	30 (°)	$\sigma_{f(\Delta\alpha_{DA},\ \Delta\alpha_1,\ \Delta\alpha_2)}$	10 (°)
$\Delta\alpha_2$	80 (°)	$\sigma_{f(\Delta\beta_{DA},\ \Delta\beta_1,\ \Delta\beta_2)}$	10 (°)
β_0	180 (°)	$\sigma_{f(r_{MA},\ r_{m,1},\ r_{m,2})}$	0.10 (Å)
$\Delta\beta_1$	70 (°)	$\sigma_{f(r_{LL},\ r_{l,1},\ r_{l,2})}$	0.10 (Å)
$\Delta\beta_2$	80 (°)	$\sigma_{\varepsilon(r,\ r_{clash})}$	0.10 (Å)

Switching Between Scoring Functions in GOLD

The GOLD code was restructured, and GOLD routines specific to scoring a docking solution were separated out into dynamically loaded libraries (dll). This allows straightforward implementation of other scoring schemes and provides a mechanism for switching between them. The Chemscore function (see below) was coded up as a scoring function dll for GOLD. The Chemscore function is described in detail by Eldridge et al.[3,12] However, because we have added terms to the original Chemscore function and modified existing terms, we describe the functional form in some detail. The parameters used are given in Table II.

The Original Chemscore Function

The Chemscore function described by Eldridge et al.[12] estimates the free energy of binding of a ligand to a protein as follows:

$$\Delta G_{binding} = \Delta G_o + \Delta G_{hbond} S_{hbond} + \Delta G_{metal} S_{metal} + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} H_{rot} \quad (2)$$

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; the functional form of this term is given by Eldridge et al.[12] The ΔG terms are coefficients derived from a multiple linear regression analysis on a training set of 82 protein–ligand complexes from the PDB (see Table II). The S_{hbond}, S_{metal}, and S_{lipo} terms all make use of a block function, f, of the following shape:

$$f(x, x_1, x_2) = \begin{cases} 1 & x \le x_1 \\ (x_2 - x)/(x_2 - x_1) & x_1 < x \le x_2 \\ 0 & x > x_2, \end{cases} \quad (3)$$

where x is a running variable, and x_1 and x_2 are constants controlling the fall-off of f. The hydrogen-bond term is calculated for each complementary combination of donor (D) and acceptor (A) on ligand and protein. It consists of a distance and an angle-dependent part:

$$S_{hbond} = \sum_{DA} f(\Delta r_{DA}, \Delta r_1, \Delta r_2) f(\Delta\alpha_{DA}, \Delta\alpha_1, \Delta\alpha_2), \quad (4)$$

with $\Delta r_{DA} = |r_{DA} - r_o|$ and $\Delta\alpha_{DA} = |\alpha_{DA} - \alpha_o|$, where r_{DA} is the H···A distance, and α_{DA} the D-H···A angle for a given donor–acceptor pair. r_o and α_o are the ideal hydrogen-bond distance and angle, respectively. Δr_1, Δr_2, $\Delta\alpha_1$ and $\Delta\alpha_2$ are constants that control the deviation from the ideal hydrogen-bond distance and angle (see Table II).

The metal term is calculated for each combination of metal (M) and acceptor (A) on ligand and protein. It only has a distance dependency:

$$S_{metal} = \sum_{MA} f(r_{MA}, r_{m,1}, r_{m,2}), \quad (5)$$

where r_{MA} is the distance between a given metal–acceptor pair. $r_{m,1}$ and $r_{m,2}$ are constants controlling the range of metal–acceptor interactions; in the original Chemscore implementation Eldridge et al. used $r_{m,1}$ = 2.2 Å and $r_{m,1}$ = 2.6 Å. Here we reoptimized these parameters (see below) and used the values in Table II, except where stated otherwise.

The lipophilic term has the same functional form as the metal term, but is much longer range:

$$S_{lipo} = \sum_{LL} f(r_{LL}, r_{l,1}, r_{l,2}). \quad (6)$$

The summation here is over all pairs of lipophilic atoms in protein and ligand. r_{LL} is the distance between protein and ligand atom for a given pair of lipophilic atoms. $r_{l,1}$ and $r_{l,2}$ are constants controlling the range of lipophilic interactions (see Table II).

The Chemscore Function Used for Docking

When the Chemscore function was adapted for docking by Baxter et al.,[3] a protein–ligand clash–energy term, E_{clash}, and a ligand–internal–energy term, E_{int}, were added to the regression-based part. Baxter et al. also added a term to penalize a docking solution moving outside a user-defined box, but because the Goldscore function does not contain such a term, we considered this unnecessary. Because GOLD has a mechanism for dealing with covalently bound ligands, we extended the Chemscore function to include a covalent energy term, E_{cov}. As a result, the form of the Chemscore function we use for protein–ligand docking is as follows:

$$\Delta G'_{binding} = \Delta G_{binding} + E_{clash} + E_{int} + E_{cov}. \quad (7)$$

The clash term is summed over all non-hydrogen protein–ligand atom pairs:

$$E_{clash} = \sum \varepsilon_{clash}(r, r_{clash}), \quad (8)$$

where r is the distance between a protein–ligand atom pair and r_{clash} is the clash distance for that pair (see Table II). The clash energy for each atom pair depends on the nature of the protein and ligand atom; it is zero for $r > r_{clash}$, and for $r \le r_{clash}$:

$$\varepsilon_{clash}(r, r_{clash}) = \begin{cases} (20/\Delta G_{hbond}) \cdot (r_{clash} - r)/r_{clash} \\ \quad \{\text{donor–acceptor pairs}\} & (9a) \\ (20/\Delta G_{metal}) \cdot (r_{clash} - r)/r_{clash} \\ \quad \{\text{metal–acceptor pairs}\} & (9b) \\ 1 + 4 \cdot (r_{clash} - r)/r_{clash} \\ \quad \{\text{all other non-H pairs}\} & (9c) \end{cases}$$

The internal energy of the ligand is the sum of a torsional term and a clash term. The latter is calculated analogously to the protein–ligand clash energy, but only for ligand atoms that are separated by at least four bonds. The torsional term reported by Baxter et al. is a summation over the ligand rotatable bonds (*RB*). Because GOLD also flips ring corners, which affects ring torsion angles, our implementation of the ligand torsional energy also includes a summation over free ring corners (*RC*), that is,

$$E_{tors} = \sum_{RB} \varepsilon_{tors}(\theta_{RB}) + \sum_{RC} \sum_{RCB} \varepsilon_{tors}(\theta_{RCB}). \quad (10)$$

The second summation in the right-hand term is over the ring bonds *RCB* affected by the ring flipping of ring corner *RC*. The functional form of $\varepsilon_{tors}(\theta)$ is given by Baxter et al.[3]

The covalent energy term only applies to ligands bound covalently to the protein. It consists of a torsional part and a bond-angle part:

$$E_{cov} = \sum_{CB} \varepsilon_{tors}(\theta_{CB}) + C_{cov} \sum_{BA} k_{BA}(\varphi_{BA} - \varphi_{o,BA})^2. \quad (11)$$

The first summation is over all torsion angles θ_{CB} involved in the covalent linkage; the second summation is over the covalent bond angles φ_{BA} around the covalent linkage. The force constants k_{BA} and the ideal bond angles $\varphi_{o,BA}$ are taken from GOLD.[5] C_{cov} is a constant used to balance the covalent bond term against the rest of the Chemscore function (see Table II). To prevent unlikely hydrogen-bond geometries, Baxter et al. modified the hydrogen-bond term to include a third term, depending on the R-A···H(D) angle, β (R being an atom attached to the acceptor, A), analogous to the r and α-dependent terms, that is,

$$S_{hbond} = \sum_{DA} [f(\Delta r_{DA}, \Delta r_1, \Delta r_2) \times f(\Delta\alpha_{DA}, \Delta\alpha_1, \Delta\alpha_2) f(\Delta\beta_{DA}, \Delta\beta_1, \Delta\beta_2)], \quad (12)$$

with $\Delta\beta_{DA} = |\beta_{DA} - \beta_o|$.

In their paper, Baxter et al. state that they used β_o = 140°, $\Delta\beta_1$ = 30°, and $\Delta\beta_2$ = 40°. This would mean that, for β = 180°, S_{hbond} = 0. We believe that this is not the

intention of the authors; therefore, here we use $\beta_o = 180°$, $\Delta\beta_1 = 70°$, and $\Delta\beta_2 = 80°$ (see Table II).

It is unclear how Baxter et al. treat acceptors that have more than one R group (e.g., OH acceptors or ring nitrogen acceptors), and hence, have more than one β-value. Here, we simply multiply the values of f for each R group. Finally, to prevent unrealistic hydrogen-bond geometries, each donor hydrogen can only form one hydrogen bond.[3]

Validation of Chemscore Implementation

The CCDC/Astex validation set contains 64 complexes from the original Chemscore set. For these complexes, we can compare our values for the different terms in the Chemscore function with those obtained by Eldridge et al.[12] When such a comparison is made, we need to bear in mind that Eldridge et al. minimized ligand and protein (under constraints) in the Discover force field, modifying hydrogen-bond positions manually, if necessary. Also, when the original Chemscore function was derived, the water molecules were left in the binding sites; most water molecules were left out of the complexes in our validation set. This means that we will not be able to reproduce the values reported by Eldridge et al. exactly, as was also noted by Sandretto and colleagues (personal communication, August 2001). In an attempt to reproduce the values reported by Eldridge et al. for the various Chemscore terms, we used the following approach:

1. We ran GOLD in "local optimization" mode, starting from the experimental binding mode, to optimize the positions of the protein and ligand OH and NH_3^+ hydrogen atoms; the *docking version* of the Chemscore function [Eq. (7)] was used here to ensure realistic hydrogen-bond geometries.
2. For the obtained binding mode, we calculated the terms in the *original* Chemscore function [Eq. (2)] (using the original parameters) and compared them with those given by Eldridge et al.

In Figure 1, we plotted the values of $\Delta G_{binding}$, calculated using the approach above, against the $\Delta G_{binding}$ values given by Eldridge et al. (minus the contribution due to interactions of water molecules). As expected, the correlation is not perfect ($R^2 = 0.84$), but is quite good in comparison with that reported by Sandretto et al. [they reported $R^2 = 0.72$ (personal communication, August 2001)], perhaps because we have optimized the positions of the OH and NH_3^+ hydrogen atoms, and we have accounted for the absence of water molecules in our test set.

If we compare the individual Chemscore terms we calculated with those reported by Eldridge et al., the hydrogen bonding and metal terms show the poorest correlations ($R^2 = 0.86$ and $R^2 = 0.80$, respectively). The metal term is very short-ranged and is sensitive to small shifts in protein and ligand. The hydrogen-bond term is also short-ranged and, additionally, directional; hence, this term is also likely to be sensitive to small changes in the positions of ligand and protein atoms, and on the orientation of ligand and protein OH and NH_3^+ groups. The lipophilic term is much longer range and not directional, and, as a result, the correlation between our values and those reported by Eldridge is high for this term ($R^2 = 0.99$); the small differences observed are probably due to rounding errors, particularly because this term is calculated on a grid (see Baxter et al.[3]). The term to penalize the freezing of ligand rotatable bonds (H_{rot}) is independent of the positions of protein and ligand atoms and should, in theory, be reproducible exactly. The correlation for this term is good ($R^2 = 0.99$). Breakdowns of the Chemscore terms, together with the corresponding protein files (with optimized OH and NH_3^+ groups), are available as supplementary material for six complexes from the "clean list."

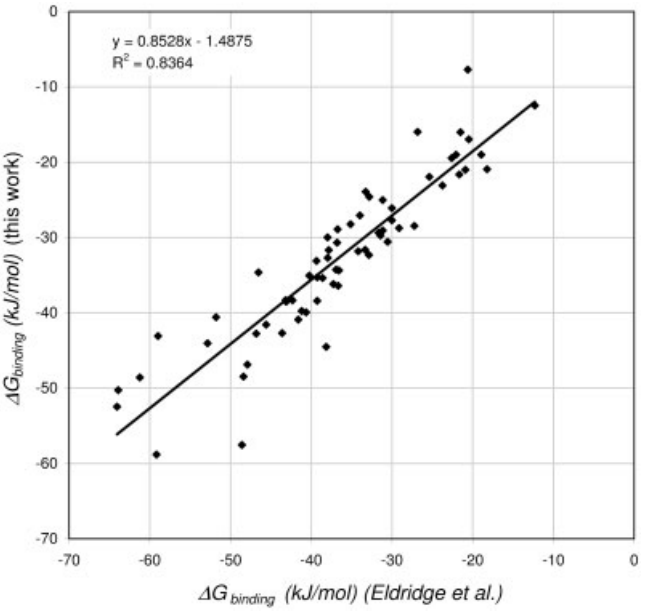


Fig. 1. Calculated $\Delta G_{binding}$ values for the Chemscore function as implemented in this work [Eq. (2)] against the calculated $\Delta G_{binding}$ values reported by Eldridge et al.[12] corrected for the interactions with water molecules.

Our implementation of the Chemscore function was also tested on four of the structure files (PDB entries 1ebg, 1fig, 1hvr, and 2tmn) used to derive the original Chemscore function by Eldridge et al. For these complexes, we observed relatively large discrepancies between our values and those reported by Eldridge et al., but when we used the exact structure files they used in their analysis, our Chemscore implementation reproduced the reported values almost exactly (for $\Delta G_{binding}$, $R^2 = 0.99$). The reasons we observed large discrepancies when we used the raw PDB files vary. For 1ebg, for example, Eldridge et al. used a different binding site for the ligand (interacting with a different protein chain); in the 2tmn complex, the zinc ion in the binding site had moved considerably during the minimization process performed by Eldridge et al.

Optimizing the Chemscore Function

Based on the results presented above, we are confident that our implementation of the Chemscore function is correct. However, because Eldridge et al. performed an

optimization of the complexes before deriving the Chemscore function, the function may not be optimal for unoptimized protein structures taken directly from the PDB. We found that GOLD's ability to predict the binding mode is particularly sensitive to the parameters in the metal term, $r_{m,1}$ and $r_{m,2}$. In the original Chemscore function, $r_{m,1}$ = 2.2 Å and $r_{m,2}$ = 2.6 Å, which makes this term extremely short range (this is probably a result of the poor description of interactions to metal ions in the force field used by Eldridge et al. to optimize the complexes). As a result, small errors in the positions of the atoms in the vicinity of the protein metal atom could easily cause a key interaction with the metal to be missed as a result of a high clash term. Therefore, the metal term was optimized against the "clean list" to $r_{m,1}$ = 2.6 Å and $r_{m,2}$ = 3.0 Å. Varying the parameters in the other terms of the Chemscore function from those used by Baxter et al. had no significant effect on performance; these parameters were therefore left unchanged.

We also introduced smoothing functions to take into account the experimental uncertainties in the protein atom positions and to make the Chemscore hypersurface less rugged. To do this, some of the terms in the Chemscore function are convoluted with a Gaussian smearing function, such that

$$y'(x) = \int_{-\infty}^{+\infty} y(x-u)g(u,\sigma)du, \qquad (13)$$

where $g(x,\sigma)$ is a normalized Gaussian smearing function with a "standard deviation" σ, $y(x)$ is the original Chemscore term, and $y'(x)$ is the smoothed Chemscore term. All f functions [see Eq. (3)] and the general clash term [Eq. (9c)] are smoothed in this way. After some optimization against the "clean list," σ values of 0.1 Å were used for all distance-dependent terms and 10° for all angle-dependent terms (see Table II).

Annealing the Chemscore Function

To promote diversity in the docking solutions, in analogy with Baxter et al.,[3] we used more relaxed hydrogen-bond parameters (Δr_2 = 1.15 Å and $\Delta\alpha_2$ = 110°) at the start of a docking run. After 75% of the GA run, the final, more restrictive parameters (see Table II) are used to focus on binding modes with good hydrogen bonds.

RESULTS AND DISCUSSION

Exhaustive Docking Runs

Figure 2 shows the cumulative percentage of complexes, as a function of the RMSD of the GOLD solution nearest to the experimental binding mode, for *exhaustive* docking runs. It is clear that, for the "clean list," the Goldscore function is significantly better at producing dockings close to the experimental binding mode (regardless of their rank) than the Chemscore function. This difference is absent in the "drug-like list" and the "fragment list," indicating that the search algorithm has a sampling problem when the Chemscore function is used. We suspect that, even after smoothing some of the terms, the Chemscore function is more "rugged" than the Goldscore function (see below). This sampling problem is reflected in the success rates: Figure 3 shows the cumulative fraction of complexes, as a function of the RMSD of the top-ranked GOLD solution. Although the Goldscore and Chemscore function perform similarly for the "drug-like list" and the "fragment list," Goldscore appears to perform better than Chemscore for the "clean list" (which contains larger ligands with more degrees of freedom).

Table III lists the success rates for exhaustive docking runs (see Table I) with the Goldscore function, the literature version of the Chemscore function, and our optimized version of the Chemscore function. The optimized version of the Chemscore function performs significantly better than the literature version but still gives significantly worse results than those obtained by Baxter et al. For an exhaustive search protocol, they report a success rate of 84% for a list of 70 complexes (for this subset, the success rate of our Chemscore implementation is 72%).[8] A number of factors may have caused this difference in success rate:

1. Baxter et al. minimized the ligand and protein before the dockings were carried out; even though this was done under severe constraints, it will have made the experimental binding mode more favorable, and, hence, the docking easier.
2. For larger compounds, when using the Chemscore function, the GOLD search algorithm does not always find the global optimum (see Fig. 2 and below).
3. Baxter et al. added a term to the Chemscore function that penalizes docking solutions for which the ligand center is outside a user-defined box; this will have made the search space smaller.
4. Here, we optimize protein OH and NH_3^+ orientations and ligand ring conformations during the docking; Baxter et al. optimized OH and NH_3^+ groups before the docking and used the ring conformations from the X-ray structures.
5. Baxter et al. used a Tabu search algorithm as opposed to the GA used here.

Table IV separates the complexes in the validation set into four categories: (1) complexes for which the binding mode is predicted correctly by both the Goldscore and the Chemscore function; (2) complexes that are predicted correctly only by the Goldscore function; (3) complexes that are predicted correctly only by the Chemscore function; and (4) complexes that are not predicted correctly by either scoring function. It is interesting to note that quite a large fraction of complexes is predicted correctly by only one of the two scoring functions. Although it is difficult to generalize the reasons why the Goldscore function performs better than the Chemscore function in certain cases, and vice versa in other cases, it does appear to be target-dependent. In a major study on neuraminidase, for example, we showed that the Goldscore function significantly outperforms the Chemscore function[23]; for p38 MAP kinase, on the other hand, the Chemscore function generally works better.

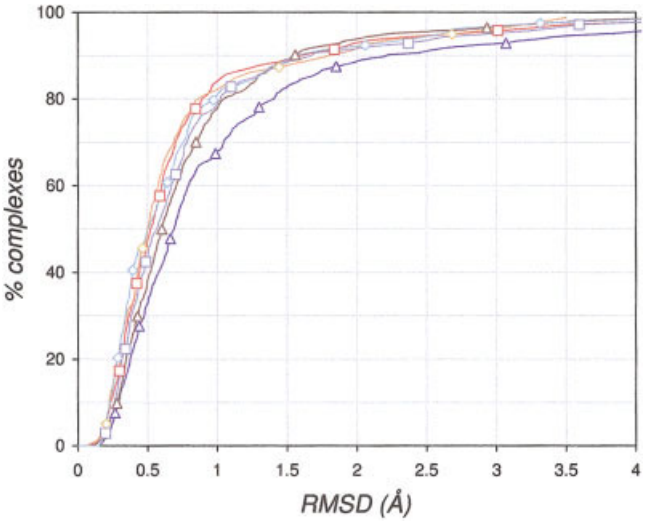


Fig. 2. Cumulative percentage of complexes as a function of the RMSD of the GOLD solution nearest to the experimental binding mode for the Goldscore function (dark-red triangles for the "clean list," red squares for the "drug-like list," and light-red diamonds for the "fragment-like list"), and for the Chemscore function (dark-blue triangles for the "clean list," blue squares for the "drug-like list," and light-blue diamonds for the "fragment-like list"). All curves are averages over three *exhaustive* GOLD runs.

Only 13% of the complexes in the "clean list" are not predicted correctly by either of the scoring functions. This indicates that there is scope for improvement if the two scoring functions were to be used in some kind of consensus manner.

Combined Docking Protocols

Table III also shows the success rates for two combined docking protocols. In the first protocol, "Goldscore-CS," the dockings produced using the Goldscore function are rescored and reranked with the optimized Chemscore function; in the second protocol, "Chemscore-GS," the dockings produced with the optimized Chemscore function are rescored and reranked with the Goldscore function; in both protocols, the Simplex algorithm is used to relax each docking in the alternative scoring function.

Interestingly, in all the cases presented in Table III, use of a second scoring function to rescore and rerank dockings gives significant improvements in success rates (3–8%), compared to straightforward docking with the first scoring function. This must be because the two combined docking protocols represent a form of consensus docking. Top-ranking but incorrect dockings produced with one scoring function score poorly in the second, ranking scoring function, hence reducing the number of false positives or "hard failures" (see below). This "consensus-docking effect" can only work if the "docking function" produces good-quality solutions for the "ranking function" to rank. The fact that the search algorithm appears to give better sampling with the Goldscore function compared to the Chemscore function (see above) probably explains why, for the "clean list" (which contains larger ligands with more degrees of freedom), "Goldscore-CS" gives better results than "Chemscore-GS."

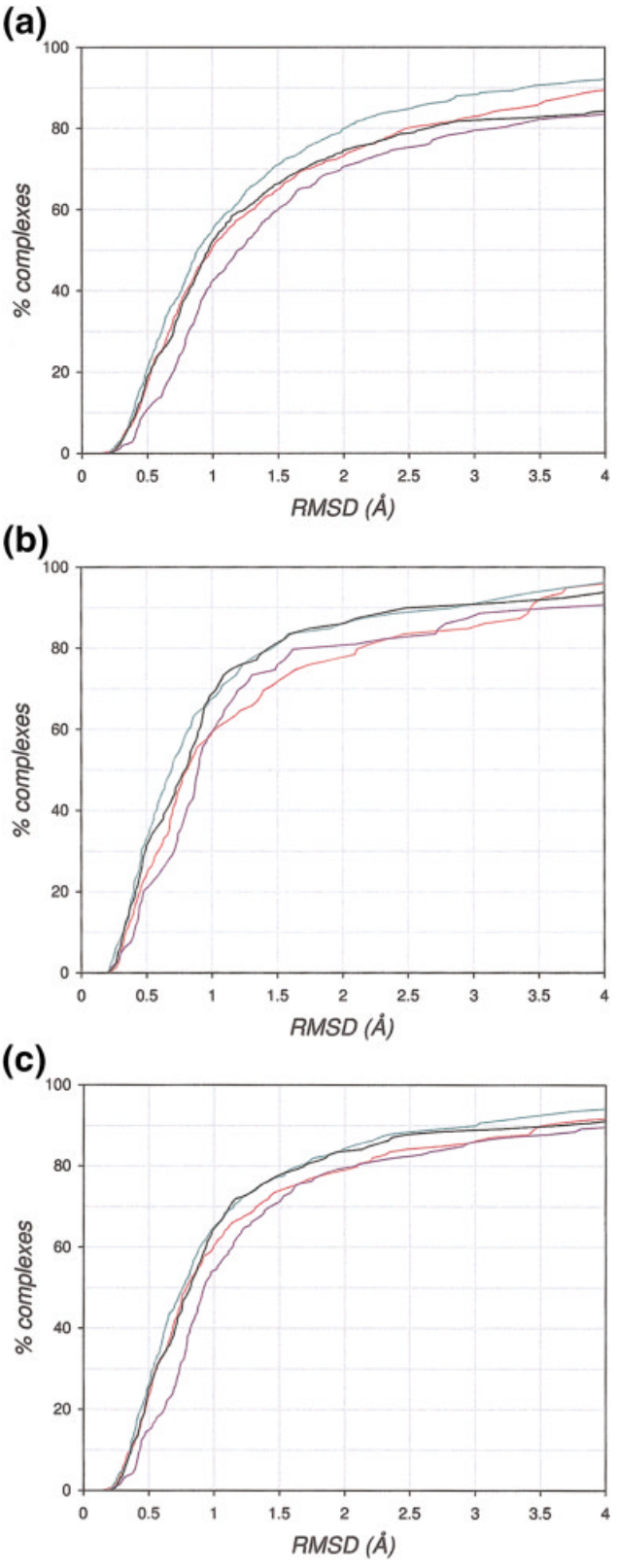


Fig. 3. Cumulative percentage of complexes as a function of the RMSD of the top-ranked GOLD solution (red for the Goldscore function, blue for the Chemscore function, green for "Goldscore-CS," and black for "Chemscore-GS") for (**a**) the "clean list," (**b**) the "drug-like list," and (**c**) the "fragment-like list." All curves are averages over three *exhaustive* GOLD runs.

TABLE III. Success Rates[a] for Exhaustive Docking Runs with Goldscore and Chemscore (All Values are Averages Over Three Runs)

	N	Goldscore	Chemscore[c]	Chemscore[d]	Goldscore-CS	Chemscore-GS
All entries	305	68.4 (1.2)[b]	63.9 (0.6)	67.2 (0.9)	74.6 (1.4)	70.6 (1.0)
Clean list	224	73.5 (1.0)	66.5 (1.2)	70.2 (1.0)	80.5 (1.8)	74.6 (1.2)
Drug-like list	147	78.9 (1.1)	75.3 (2.3)	79.9 (1.4)	84.4 (1.5)	83.5 (1.2)
Fragment list	79	78.5 (0.0)	75.1 (2.6)	81.0 (0.0)	86.5 (1.9)	86.5 (0.7)

[a]Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å RMSD of the experimental binding mode.
[b]Standard deviations are given in parentheses. These standard deviations only take into account the nondeterministic nature of the search algorithm; they do not include sampling errors, which are related to the size of the validation set (see Nissink et al.[7]). Assuming an overall success rate of 68%, this error is 2.6%, 3.1%, and 5.8% for the overall set, the "clean list," and the "fragment list," respectively.
[c]Literature Chemscore function.[3]
[d]Optimized Chemscore function (this work).

TABLE IV. Percentages of Complexes Predicted Correctly by Both the Goldscore and the Chemscore Function, Only the Goldscore Function, Only the Chemscore Function, or by Neither Scoring Function

	N	Both	Goldscore only	Chemscore only	Neither
Clean list	224	56.9 (1.4)	16.6 (1.3)	13.3 (1.3)	13.2 (1.3)
Drug-like list	139	67.6 (1.3)	11.3 (1.4)	12.2 (0.5)	8.9 (0.7)
Fragment list	79	66.2 (0.6)	12.2 (0.6)	14.8 (0.6)	6.8 (0.6)

All values are averages of the nine combinations of the three *exhaustive* Goldscore runs and the three *exhaustive* Chemscore runs. Standard deviations are given in parenthesis.

The success rates obtained with the combined docking protocols are significantly higher than those reported in previous GOLD validations. "Goldscore-CS" gives a 75% success rate on the complete validation set, an 81% success rate for the "clean list," and success rates of around 85% for the "drug-like list" and "fragment list." "Chemscore-GS" gives similar success rates for the "drug-like list" and the "fragment list," but not for the "clean list" (and the overall set); this is also clear from Figure 3.

Consensus scoring and consensus docking are not new concepts to protein–ligand docking. Consensus *scoring* is aimed at improving hit rates when libraries of compounds are screened.[24] Each compound in the library is docked and then rescored with various scoring functions, and consensus ranks are calculated. Consensus *docking*, the focus of this article, is aimed at improving the docking performance. Various approaches to consensus docking have been reported in the literature. Paul and Rognan clustered docking solutions from DOCK, FlexX, and GOLD into "consensus pairs" and obtained improved docking success rates.[25] Clark et al. observed improved docking performance when they rescored FlexX docking solutions with various scoring functions and combined the scores into a consensus score, CScore.[26] Similar results were obtained by Terp et al. when they reranked GOLD docking solutions, using their consensus scoring function Multi-Score.[27] The examples of consensus scoring that are most relevant to this work are articles by Hoffman et al.[28] and Gohlke et al.[29]; both represent examples of simple rescoring and reranking of docking solutions with a second scoring function. Hoffman et al. rescored FlexX docking solutions using a classical force field and observed significant improvements in the docking performance. Gohlke et al. also observed significant improvement in docking performance when FlexX and Dock docking solutions were rescored with the DrugScore scoring function.

Performance Versus Time

The results listed in Table III were obtained with use of the *exhaustive* search protocol (see Table I). This was done to minimize the effect of the search algorithm and to compare the performance of the different scoring functions, but the average CPU time per complex using these GA settings is in the order of 2–4 h. This is not realistic for virtual screening, in which, typically, many thousands of compounds are docked against a target binding site. It is therefore key to use search settings that are as fast as possible, with a minimal loss of docking performance. Table V lists the performance of GOLD for various GA settings (see Table I), using the Goldscore and the Chemscore function, and the two combined docking protocols.

First, it is clear that, particularly for complexes with "drug-like" or "fragment-like" ligands, the Goldscore and Chemscore function give similar results in terms of their docking accuracy. Importantly, the Chemscore function is considerably faster than the Goldscore function. For identical search settings and similar docking performance, the Chemscore function provides an increase in speed of up to three-fold compared to the Goldscore function. We believe there are two main reasons why the evaluation of the Chemscore function is faster than that of the Goldscore function:

1. The lipophilic and clash terms in the Chemscore function do not take hydrogen atoms into account. As a

TABLE V. Success Rates[a] and Average Docking Times[b] for Different GA Settings

	Goldscore		Chemscore		Goldscore-CS	Chemscore-GS
GA setting	Time	Success	Time	Success	Success	Success
Clean list (224 complexes)						
Exhaustive	248.9	73.5 (1.0)	117.2	70.2 (1.0)	80.5 (1.8)	74.6 (1.2)
GOLD default 1	24.7	72.9 (1.4)	13.2	68.9 (1.7)	76.1 (1.5)	69.8 (1.6)
GOLD default 2	9.2	70.2 (1.6)	4.4	66.3 (1.5)	73.8 (2.0)	67.5 (1.8)
GOLD default 3	5.9	66.4 (1.6)	2.6	65.3 (1.7)	69.8 (2.1)	65.9 (1.6)
GOLD default 4	2.1	63.0 (1.8)	0.74	60.9 (1.8)	65.6 (2.3)	62.0 (2.2)
GOLD library	0.43	56.0 (1.9)	0.12	52.5 (1.9)	58.0 (2.1)	52.0 (2.1)
Drug-like list (139 complexes)						
	Goldscore		Chemscore		Goldscore-CS	Chemscore-GS
	Time	Success	Time	Success	Success	Success
Exhaustive	188.2	78.9 (1.1)	86.6	79.9 (1.4)	84.4 (1.5)	83.5 (1.2)
GOLD default 1	11.6	79.1 (1.4)	5.7	78.8 (1.6)	82.9 (1.4)	79.9 (1.8)
GOLD default 2	4.9	79.5 (1.3)	2.3	77.5 (1.7)	82.9 (1.5)	79.0 (2.0)
GOLD default 3	3.2	75.3 (1.9)	1.3	77.5 (1.9)	78.1 (1.8)	78.5 (1.8)
GOLD default 4	1.1	73.4 (2.1)	0.37	74.8 (2.1)	76.1 (2.2)	77.0 (2.3)
GOLD library	0.25	71.3 (2.2)	0.07	69.4 (2.4)	74.6 (2.2)	69.8 (2.5)
Fragment list (79 complexes)						
	Goldscore		Chemscore		Goldscore-CS	Chemscore-GS
	Time	Success	Time	Success	Success	Success
Exhaustive	136.3	78.5 (0.0)	67.6	81.0 (0.0)	86.5 (1.9)	86.5 (0.7)
GOLD default 1	5.1	78.5 (1.3)	3.3	82.8 (2.1)	83.0 (2.1)	85.1 (1.7)
GOLD default 2	2.8	79.5 (1.3)	1.5	83.1 (2.0)	84.5 (2.1)	84.7 (2.0)
GOLD default 3	1.8	78.0 (2.1)	0.87	82.9 (2.3)	81.9 (2.4)	84.7 (1.8)
GOLD default 4	0.69	76.4 (2.8)	0.25	81.1 (2.5)	80.1 (2.2)	84.4 (2.6)
GOLD library	0.15	78.6 (3.9)	0.05	77.9 (2.9)	83.1 (2.7)	80.1 (2.9)

Note: All results are averages of 50 runs except those for the *exhaustive* settings, which are averages of three runs.

[a]Percentage of complexes for which the top-ranked GOLD solution is within 2.0 Å of the experimental binding mode.

[b]Single-processor CPU minutes for the complete docking of a ligand, excluding the protein initialization time (which varies between 0.5 and 3.0 min). All calculations were done on an 84-processor Linux cluster of 1GHz/PentiumIII PC's. Times for "Goldscore-CS" are identical to those for the Goldscore function. Times for "Chemscore-GS" are identical to those for the Chemscore function.

result, contrary to the Goldscore S_{vdw_ext} term, these two terms can be precalculated on grids.

2. The functional form of the ligand intramolecular energy of the Chemscore function is simpler than that of the Goldscore function. We must stress, however, that, although here we have made no attempt to do so here, it may well be possible to speed up the Goldscore function considerably.

As we observed for the *exhaustive* search settings, in nearly all cases presented in Table V, using a second scoring function to rescore and rerank dockings gives significant improvements in success rates compared to straightforward docking using the first scoring function. It is interesting to note that these improvements in success rates drop with faster search settings, because the consensus docking approaches rely on adequate sampling of the possible binding modes.

As expected, the performance of each scoring function drops with shorter search times; however, much faster search settings can be used without a dramatic loss of performance, particularly for drug-like and fragment-like ligands. For fragment-like compounds, "Chemscore-GS," combined with the *GOLD default 4* settings provides the best balance between performance and speed (84% success at 0.25 min/complex). It may appear from Table V that the *GOLD library* settings also provide reasonable success rates for fragment-type ligands. The problem is that for fragment-type compounds, an RMSD threshold of 2.0 Å to define success is not very appropriate; for small ligands, we need to more critically assess whether the binding mode is predicted correctly, and, although at 2.0 Å RMSD, the performance of the *GOLD library* settings and the *GOLD default 4* settings is similar, at lower RMSD thresholds the latter settings give superior results.

For drug-like compounds, "Chemscore-GS" combined with the *GOLD default 3* or even the *GOLD default 4* settings, seems to be optimal (~78% success at 0.37–1.3 min/complex). For larger compounds in the "clean list," "Goldscore-CS" in combination with the *GOLD default 2* settings appears the most appropriate choice (74% success at 9.2 min/complex).

The results in Table V indicate that, in terms of speed and docking accuracy, GOLD is an excellent tool for docking and virtual screening. Drug-like compounds can routinely be docked in 0.5–1.0 min/compound (i.e., ~2000 compounds/day on a single processor), and the docking accuracy is good (~78% at 2.0 Å RMSD, ~68% at 1.5 Å RMSD, ~55% at 1.0 Å RMSD).

GOLD is often perceived to be too slow for large-scale virtual screening. Recently, Lyne related the suitability of a docking program for large-scale virtual screening to its ability to screen 100,000 compounds on eight processors in a few days.[30] On this basis, Lyne considered GOLD's suitability for large-scale virtual screening to be "low." Screening 100,000 compounds on eight of the processors we used in this work would take approximately 6 days. But computing power is becoming very cheap, and large-scale virtual screening is typically done on Linux clusters that consist of more than eight processors. On the 84-processor cluster used in this work, we can screen approximately 1.2 million compounds/week using GOLD. Also, processors have increasingly become faster. For example, 2.4 GHz Xeon processors are 2.5 times faster for GOLD applications than the processors we used here.

It is interesting to note that even with the *GOLD library* settings, which represent a trivial amount of searching (see Table I), 52–58% (depending on the scoring protocol) of the "clean list" can be predicted within 2.0 Å of the experimental binding mode. We already mentioned that, for smaller ligands, the 2.0 Å cutoff is not very useful, but even at a 1.0 Å cutoff, 30–37% of the complexes in the "clean list" can be predicted successfully with the *GOLD library* settings. We believe this is not just the case for our validation set, but that it is quite typical for validation sets used for testing docking programs. Hence, we feel that the challenge for workers developing docking programs or scoring functions is to push success rates beyond the 50–60% mark (at a 2.0Å cutoff), and aim to predict more of the remaining, more difficult complexes correctly.

Soft Versus Hard Failures

Verkhivker and colleagues[13] introduced the terminology "soft failures" and "hard failures" to categorize reasons why a docking program is unable to predict the binding mode of a ligand successfully. Soft failures arise when the search algorithm has not found the global optimum, whereas hard failures arise when the global optimum does not correspond to a binding mode close to the experimental binding mode. Because it is difficult to prove whether the algorithm has found the global optimum, our definitions of soft and hard failures are as follows: A soft failure occurs when the top-ranked GOLD solution is not within the chosen RMSD cutoff (usually 2.0 Å), and the score of the experimental binding mode (after local optimization) is better than that of the top-ranked GOLD solution; a hard failure occurs when the top-ranked GOLD solution is not within the RMSD cutoff, and the score of the experimental binding mode is worse than that of the top-ranked GOLD solution.

Figure 4 shows the occurrence of hard and soft failures for the four scoring protocols we describe here, and for the different GA settings. First, as expected, with faster search settings, the fraction of soft failures goes up. The fraction of hard failures stays more or less constant for different GA settings, except for the *GOLD library* settings, in which many of the soft failures are actually hard failures in the original definitions used by Verkhivker et al. Again, it is clear that, even for the *exhaustive* settings, the search algorithm does not always find the global optimum when the Chemscore function is used. Whereas docking with the Goldscore function only produces 1% soft failures for the "clean list," docking with the Chemscore function produces 6% soft failures; for the "drug-like list" and the "fragment list," this difference is much smaller. This indicates that, in principle, the Chemscore function is at least as good a scoring function as the Goldscore function, but that the search algorithm finds it more difficult to find the global optimum.

In all cases presented in Figure 4, except for the *GOLD library* settings (see above), the use of a second scoring function to rescore and rerank dockings significantly reduces the fraction of hard failures compared to straightforward docking using the first scoring function. As we discussed, reducing hard failures, or false positives, is the mechanism by which the combined docking protocols work. The number of soft failures, however, rises when the combined docking protocols are used. This is to be expected, because the "docking function" is not aimed at finding the optima of the "ranking function."

Estimation of Binding Affinities

Figure 5(a) shows a plot of the Chemscore $\Delta G_{binding}$ for the experimental binding modes, against the experimental binding energies, for the 60 complexes in the Chemscore set for which reliable binding data are available. Interestingly, the Goldscore function [Fig. 5(b)] gives an equally good correlation with affinity as the Chemscore $\Delta G_{binding}$. This is surprising, because, unlike the Chemscore function, the Goldscore function was never parameterized against binding affinities. It is important, however, to subtract the intramolecular terms from the *GOLD Fitness* [see Eq. (1)]; as far as we know, this has never been reported before in the literature (from here on, we refer to the *GOLD Fitness*, minus the intramolecular terms, as the *Goldscore*, not to be confused with the *GOLD Fitness* or the *Goldscore function*). The intramolecular terms have an arbitrary reference point and, although essential for successful docking, are meaningless when different compounds are compared. Figure 5(c) shows the dramatic loss of correlation with affinity when the intramolecular terms are not subtracted.

Figure 6 shows GOLD's ability to predict the binding energy for the various search settings, and for the different scoring protocols (Goldscore, Chemscore, "Goldscore-CS," and "Chemscore-GS"). For all scoring protocols, the correlation of the scores with experimental binding energies deteriorates with faster search settings; this highlights the importance of correctly predicting the binding modes to obtain reasonable estimates of the binding energy. What

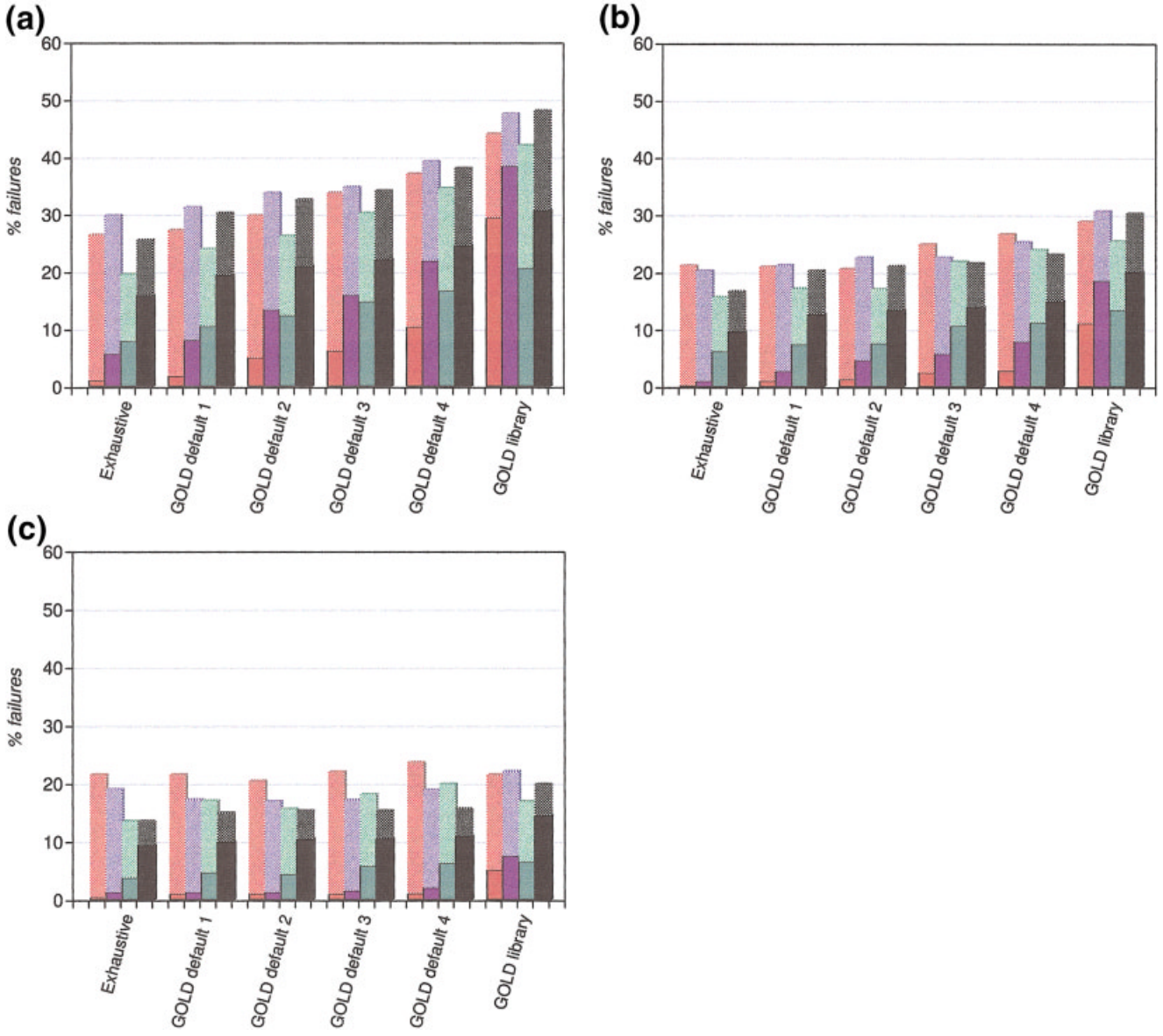
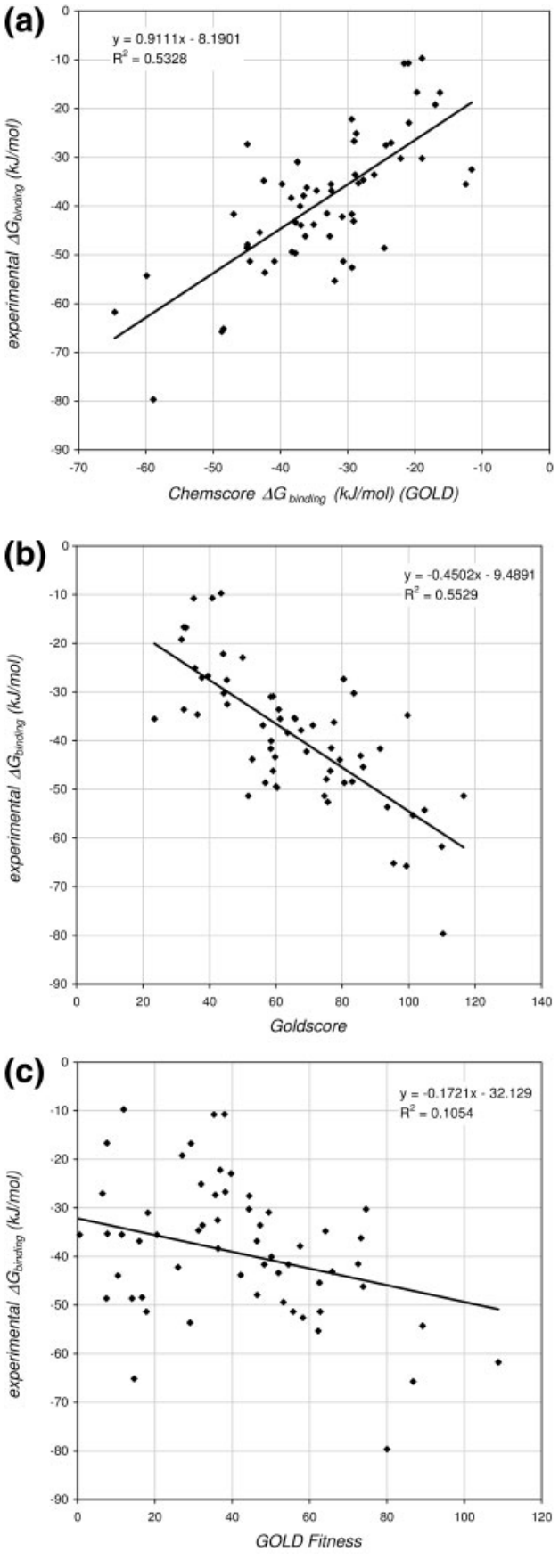


Fig. 4. Percentage soft failures (darker colors) and hard failures (lighter colors) for (**a**) the "clean list," (**b**) the "drug-like list," and (**c**) the "fragment-like list" (red for the Goldscore function, blue for the Chemscore function, green for "Goldscore-CS," and black for "Chemscore-GS"). All values are averages of 50 GOLD runs, except those for the *exhaustive* settings, which are averages of three runs.

is immediately clear from Figure 6 is that scores produced with the *GOLD library* settings show no correlation with the experimental binding affinity (for any of the docking protocols).

Although the Chemscore function (blue columns) performs as well as the Goldscore function for the X-ray binding modes and the *exhaustive* search settings, its performance drops rapidly for faster search settings. A similar, rapid drop in performance for faster search protocols was observed by Baxter et al.[8] For faster search settings, the Goldscore function (red columns) performs much better than the Chemscore function; the performance for the *GOLD default 3* settings is nearly as good as that for the *exhaustive* settings, and the performance of the *GOLD default 4* settings is only slightly worse. Although the combined docking protocols produce more accurate dockings (see above), the final scores obtained with these protocols do not correlate better with the experimental binding affinities. In fact, typically, the performance of the "Goldscore-CS" (green columns) and the "Chemscore-GS" (black columns) protocols is similar to that of the Chemscore function. It has to be pointed out that the s values we obtained here correspond to an error of nearly two orders of magnitude in affinity. Although this is state of the art for fast-scoring functions, it is clearly not ideal, particularly for the lead-optimization stages of a drug discovery project, which stresses the ongoing need to improve the current scoring functions.

Many of the compounds in the Chemscore test set used in the above analysis are large and un-drug-like, making



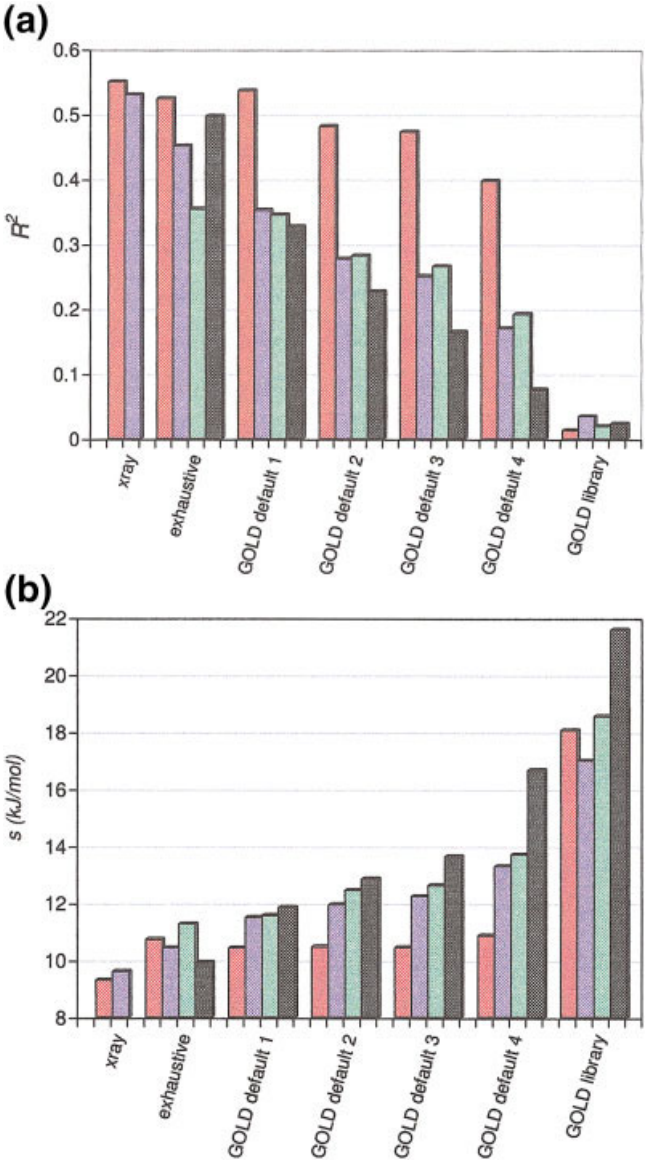


Fig. 6. (**a**) Correlation coefficients R^2 and (**b**) standard deviations *s* of predicted binding affinities for the X-ray binding mode and dockings with the various search settings (red for the Goldscore function, blue for the Chemscore function, green for "Goldscore-CS," and black for "Chemscore-GS"). All values are averages of 50 GOLD runs, except those for the *exhaustive* settings, which are averages of three runs; for the Chemscore and "Goldscore-CS" dockings, the predicted binding affinity was calculated as $\Delta G_{binding} = 0.9111 \cdot \Delta G_{binding}$ (*Chemscore*) $- 8.1901$ [see Fig. 5(a)]. For the Goldscore and "Chemscore-GS" dockings, the predicted binding affinity was calculated as $\Delta G_{binding} = -0.4502 \cdot$ *Goldscore* $- 9.4891$ [see Fig. 5(b)].

the search problem more difficult than it would be in a typical drug discovery project. Hence, we ran the same analysis on the complexes in the Chemscore set for which

Fig. 5. Experimental binding energy against (**a**) the Chemscore $\Delta G_{binding}$ as implemented in GOLD, (**b**) the *Goldscore* (i.e., the *GOLD Fitness*, minus intramolecular terms), and (**c**) the GOLD Fitness. Chemscore $\Delta G_{binding}$, *Goldscore*, and *GOLD Fitness* are all calculated by running GOLD in *local optimization* mode (without Simplex).

the ligands are drug-like. Unfortunately, the Chemscore set only contains 36 complexes that fall into this category, so the results are not as statistically significant as those presented in Figure 6. But the results do indicate that, for drug-like compounds, the errors in the predicted binding energies are smaller for faster search settings than those for the complete Chemscore set. In fact, for the *GOLD default 4* settings, $s \approx 9$–10 kJ/mol for all four scoring protocols. These results indicate that the performance of the four docking protocols, as presented in Figure 6, is not necessarily representative for their performance in virtual screening applications, because the compounds screened typically are relatively small and drug-like. Additionally, in virtual screening, the objective is to identify potential binders in a database of mainly inactive compounds, rather than to accurately predict the relative binding affinities of a set of known binders. We will address the performance of the four docking protocols in virtual screening applications in a separate study.

CONCLUSIONS

In this article, we have described the implementation of the Chemscore function as a scoring function for GOLD. The performance of the Chemscore function, in terms of docking accuracy, its ability to predict binding affinities, and its speed, were compared to the original Goldscore function. The Chemscore function gives equally good docking results as the Goldscore function for "drug-like" and "fragment-like" ligands, but for larger ligands, Goldscore gives superior results. This, together with several other observations, indicates that the search algorithm has a sampling problem when the Chemscore function is used for docking larger compounds, probably because its "energy landscape" is more "frustrated" than that of the Goldscore function. A major advantage of using the Chemscore function is that the dockings are up to three times faster when compared with the Goldscore function.

The performance of two combined, or consensus, docking protocols that use both the Goldscore and the Chemscore function was also analyzed. In the first protocol, "Goldscore-CS," dockings produced with the Goldscore function are scored and ranked with the Chemscore function; in the second protocol, "Chemscore-GS," dockings produced with the Chemscore function are scored and ranked with the Goldscore function. Both protocols give significant improvements in docking accuracy over the use of the Goldscore or Chemscore function alone, by reducing the number of "hard docking failures."

The performance of all four docking protocols was tested as a function of the search settings (i.e., the speed of the dockings). For most virtual screening applications, "Chemscore-GS" combined with the *GOLD default 3* or even the *GOLD default 4* search settings seems optimal. This gives docking accuracies of around 78% (top-ranked GOLD solution is within 2.0 Å of the experimental binding mode) for "drug-like" compounds and 85% for "fragment-like" compounds; docking speeds are in the order of 0.25–1.3 min/compound on a single CPU. For larger ligands, when the main objective is to predict the binding mode, "Goldscore-CS" is most appropriate; if the *exhaustive* search settings are used, success rates as high as 81% can be obtained for the "clean list," but at the cost of very long search times. The *GOLD default 1* (76% at 24 min/compound) or *GOLD default 2* (74% at 9 min/compound) search settings may be more practical in this case.

In terms of the ability to predict accurate binding affinities, surprisingly, in our test set, the Goldscore function outperforms the Chemscore function, particularly for faster search settings. Even at docking speeds of around 1–2 min/compound, the Goldscore function predicts binding energies with a standard deviation of ~10.5 kJ/mol. It is key, however, when the Goldscore function is used to predict binding energies, that the ligand intramolecular terms are subtracted from the *GOLD Fitness*. The two combined docking protocols perform similarly to the Chemscore function in terms of the accuracy of the predicted binding energies.

The above recommendations are based on the performance of GOLD, averaged over a large number of complexes. It was observed, however, that the performance of the various docking protocols is target dependent. Therefore, when using GOLD in a structure-based project, our advice is to test various scoring and searching protocols, and to select those that give optimal performance.

ACKNOWLEDGMENTS

Our thanks to Jin Li and Bohdan Waszkowycz for providing us with some structures and data to validate our implementation of the Chemscore function, and to Kathryn Sandretto for sending us details of the performance of her implementation of the Chemscore function.

REFERENCES

1. Makino S, Kuntz ID. 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. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins 1998;33:367–382.
4. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
5. Jones G, Willett P, Glen RC. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol 1995;245:43–53.
6. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput-Aided Mol Des 2002;16:151–166.
7. Nissink JWM, Murray CW, Hartshorn MJ, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein–ligand interaction. Proteins 2002;49:457–471.
8. Baxter CA, Murray CW, Waszkowycz B, Li J, Sykes RA, Bone RGA, Perkins TDJ, 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.
9. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins 1999;37:228–241.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

11. Bohm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. J Comput-Aided Mol Des 1994;8: 243–256.
12. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. 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.
13. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Arthurs S, Colson AB, Freer ST, Larson V, Luty BA, Marrone T, Rose PW. Deciphering common failures in molecular docking of ligand–protein complexes. J Comput-Aided Mol Des 2000;14:731–751.
14. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 2000;44:235–249.
15. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46:3–26.
16. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002;45:2615–2623.
17. Blundell TL, Jhoti H, Abell C. High-throughput crystallography for lead discovery in drug design. Nat Rev Drug Discov 2002;1:45–54.
18. Nienaber VL, Richardson PL, Klighofer V, Bouska JJ, Giranda VL, Greer J. Discovering novel ligands for macromolecules using X-ray crystallographic screening. Nat Biotechnol 2000;18:1105–1108.
19. Hajduk PJ, Meadows RP, Fesik SW. NMR-based screening in drug discovery. Q Rev Biophys 1999;32:211–240.
20. Fejzo J, Lepre CA, Peng JW, Bemis GW, Ajay, Murcko MA, Moore JM. The SHAPES strategy: An NMR-based approach for lead generation in drug discovery. Chem Biol 1999;6:755–769.
21. Boehm HJ, 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.
22. Carr R, Jhoti H. Structure-based screening of low-affinity compounds. Drug Discov Today 2002;7:522–527.
23. Birch L, Murray CW, Hartshorn MJ, Tickle I, Verdonk ML. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. J Comput-Aided Mol Des 2002;16:855–869.
24. Charifson PS, Corkery JJ, Murcko MA, Walters WP. 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.
25. Paul N, Rognan D. ConsDock: A new program for the consensus analysis of protein–ligand interactions. Proteins 2002;47:521–533.
26. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20:281–295.
27. Terp GE, Johansen BN, Christensen IT, Jorgensen FS. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein–ligand binding affinities. J Med Chem 2001;44:2333–2343.
28. Hoffmann D, Kramer B, Washio T, Steinmetzer T, Rarey M, Lengauer T. Two-stage method for protein–ligand docking. J Med Chem 1999;42:4422–4433.
29. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol 2000;295: 337–356.
30. Lyne PD. Structure-based virtual screening: An overview. Drug Discov Today 2002;7:1047–1055.

SUPPLEMENTARY MATERIAL

TABLE SI. Breakdown of the GOLD Fitness for a Selection of Complexes in the "Clean list"

PDB entry	GOLD Fitness	Goldscore	S_{hb_ext}	S_{vdw_ext}	S_{vdw_int}	S_{cov}
1abe	50.36	58.50	26.42	32.09	−8.14	0.00
1aec	39.68	66.38	9.10	57.28	−18.74	−7.95
1c83	29.85	72.61	25.74	46.87	−42.76	0.00
1hiv	97.71	120.29	11.93	108.36	−22.58	0.00
2tmn	50.11	63.34	27.57	35.77	−13.23	0.00
4dfr	63.71	102.60	40.29	62.31	−38.89	0.00

TABLE SII. Breakdown of the Chemscore Terms for a Selection of Complexes in the "Clean List," Using the Original Chemscore Parameters in Baxter et al.[3]

PDB entry	$\Delta G'_{binding}$	$\Delta G_{binding}$	S_{hbond}	S_{metal}	S_{lipo}	E_{clash}	E_{int}	H_{rot}	E_{cov}
1abe	−23.68	−23.88	5.87	0.00	60.83	0.00	0.21	3.25	0.00
1aec	−15.86	−21.20	5.86	0.00	116.73	2.35	1.08	6.83	1.90
1c83	−35.54	−40.36	6.35	0.00	116.85	1.02	3.81	0.00	0.00
1hiv	−43.15	−49.29	6.81	0.00	315.34	0.00	6.14	6.18	0.00
2tmn	−22.68	−24.25	4.17	0.78	56.58	0.00	1.58	2.53	0.00
4dfr	−33.77	−36.23	7.50	0.00	142.29	0.00	2.47	4.27	0.00

TABLE SIII. Breakdown of the Chemscore Terms for a Selection of Complexes in the "Clean List," Using the Chemscore Parameters in this Work (see Table II)

PDB entry	$\Delta G'_{binding}$	$\Delta G_{binding}$	S_{hbond}	S_{metal}	S_{lipo}	E_{clash}	E_{int}	H_{rot}	E_{cov}
1abe	−21.87	−22.28	5.39	0.00	60.93	0.21	0.21	3.25	0.00
1aec	−15.12	−20.29	5.58	0.00	116.86	2.19	1.08	6.83	1.90
1c83	−32.87	−38.09	5.67	0.00	116.87	1.43	3.79	0.00	0.00
1hiv	−43.83	−49.69	6.92	0.00	315.54	0.18	6.11	6.18	0.00
2tmn	−27.21	−29.45	4.07	1.70	56.57	0.67	1.58	2.53	0.00
4dfr	−32.61	−34.66	7.02	0.00	142.35	0.00	2.04	4.27	0.00