Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions

ALFONSO T. GARCÍA-SOSA, CSABA HETÉNYI, UKO MARAN

Institute of Chemistry, University of Tartu, Jakobi 2, Tartu 51014, Estonia

Abstract: A dataset of protein-drug complexes with experimental binding energy and crystal structure were analyzed and the performance of different docking engines and scoring functions (as well as components of these) for predicting the free energy of binding and several ligand efficiency indices were compared. The aim was not to evaluate the best docking method, but to determine the effect of different efficiency indices on the experimental and predicted free energy. Some ligand efficiency indices, such as $\Delta G/W$ (Wiener index), $\Delta G/NoC$ (number of carbons), and $\Delta G/P$ (partition coefficient), improve the correlation between experimental and calculated values. This effect was shown to be valid across the different scoring functions and docking programs. It also removes the common bias of scoring functions in favor of larger ligands. For all scoring functions, the efficiency indices effectively normalize the free energy derived indices, to give values closer to experiment. Compound collection filtering can be done prior or after docking, using pharmacokinetic as well as pharmacodynamic profiles. Achieving these better correlations with experiment can improve the ability of docking scoring functions to predict active molecules in virtual screening.

© 2009 Wiley Periodicals, Inc.   J Comput Chem 31: 174–184, 2010

Key words: virtual screening; scoring function; drug design; docking; free energy of binding

Introduction

Many drugs have been developed with the use of structure-based drug design and molecular docking.[1–5] When used correctly, docking can be an invaluable tool for drug discovery and design. Commonly, docking is used as a complement to other techniques such as high-throughput screening (HTS). However, as an example, Pierce et al. show that it can also be the primary technique, predicting 4 kinase inhibitors with a 14-fold increase in enrichment over HTS.[6] The active molecules' binding modes predicted by docking were experimentally confirmed by X-ray crystallography.

Docking scoring functions perform generally well for predicting protein-ligand binding modes,[1–5] although they are less accurate for predicting binding free energy.[1–5,7] Docking programs employ at least one scoring function for calculating the fit or energy of a protein-ligand association. Scoring functions are usually derived from atomic parameters generated from empirical or knowledge-based approximations to the experimental binding energy of protein-ligand complexes. Most scoring functions are additive in nature, in the sense that the more functional groups a ligand has, the more interactions it can have with the protein and the greater the intermolecular energy is thus calculated. In the case of polar functional groups, this would normally be offset by higher desolvation energies, which are unfavorable to the overall binding free energy. However, these desolvation energies, if included in the scoring function or docking program at all, do not tend to reflect the real trends, and so the scoring functions end up overestimating the binding energy for larger ligands at the expense of smaller ligands.[1] A similar situation arises for large hydrophobic ligands because the larger the molecule, the more van der Waals contacts are calculated. Again, large molecules would also incur in entropy penalties when binding and even if some scoring functions attempt to estimate this entropy loss by a measure of the number of rotatable bonds of a ligand, they are not accurate and end-up still favoring larger molecules. The proper calculation of entropies of binding is also a complex issue for scoring functions, unlikely to be solved by simple rotatable bond counts.[8]

An inaccuracy of only 1–2 kcal/mol represents already a difference of one or more orders of magnitude in the calculated affinities of proteins for ligands. However, even within this level of inaccuracy, docking should be able to classify ligands as having milli-, micro-, or nanomolar affinity in order to have predictive ability (a difference of 3 orders of magnitude is around 4 kcal/mol). In drug discovery and design, accuracy is arguably a more critical value to achieve than extreme precision. In other words, accurate relative ranking of diverse and unrelated active and inactive compounds is more sought after than less accurate but precise absolute binding energies. In any case, there is a need for computed values with higher accuracy that can be compared with experimental data. This would reduce the number of false positive and false negatives that a virtual screen can produce. The development of better scoring functions and docking methods is an active field of research,[1,2] with improvements likely to come from better descriptions and parameterizations of binding,[9–12] solvation interactions,[13–17] as well as flexibility[18–20] and entropy effects.[8,21]

Another method of improving the result from docking experiments is postprocessing the results, such as combining the result of several scoring functions, called "consensus scoring,"[7,22] or rescoring the energy for docked poses with a different method, such as molecular mechanics Poisson-Boltzmann (MM/PBSA)[23] or generalized Born/Surface Area (MM/GBSA).[24,25] Comparisons between scoring functions and related challenges have been performed elsewhere,[26,27] and it is not the objective of this article.

Recently, ligand efficiency indices (E.I. = ΔG/Measure, where ΔG is the binding free energy) have been proposed as a method to normalize the experimental,[28–31] as well as computational binding free energies of ligands.[32–35] An efficiency index measure can be any molecular measure of comparison between ligands and can be related to the molecular size such as molecular weight (MW), number of heavy atoms (NHA), number of carbons (NoC), or molecular or polar surface area. They can also be related to the solubility and permeability of a ligand by incorporating the logarithm of the octanol-water partition coefficient log P.[34] They can provide a measure of how efficiently a ligand binds to a biomolecule, even being able to determine the compounds that may disrupt a protein–protein interaction if their number of heavy atoms efficiency index, ΔG/NHA, is deeper than -0.24 kcal/molNHA.[35] The reason for this is that those small molecules have a higher efficiency of binding per heavy atom than the protein or peptide they displace, even with a surface area as low as half that of the peptide or protein.[35] There is a well-known tendency of lead molecules to increase in size and lipophilicity during optimization in search of higher affinity.[36] But this increase in lipophilicity can also carry more risks in associated side-effects and toxicity.[36] The related measure of pIC50 − cLogP has also been introduced to try to define the lipophilic space available to drug candidates.[36]

In this work, we explore how different docking programs and scoring functions can correlate with experimental values, both for free energy of binding, as well as to five different efficiency indices. These efficiency indices are free energy of binding/molecular weight (ΔG/MW), free energy of binding/number of heavy atoms (ΔG/NHA), free energy of binding/number of carbons (ΔG/NoC), logarithm of −free energy of binding/partition coefficient (log($-\Delta G/P$)), and free energy of binding/Wiener index (ΔG/W).[33,34] Achieving better correlations of scoring functions with experimental values can increase the accuracy of scoring functions, and therefore the reliability of docking programs to predict active molecules in screening procedures.

Results

It is important to use drug compounds as test systems, because their complexity as compared with standard compounds can be challenging for scoring functions. Twenty-six protein-drug complexes with known experimental free energy of binding were obtained from comparing the PDBbind and DrugBank databases, yielding a wide variety of drugs with different shapes, sizes and chemical features. These complexes are given in the Supporting Information Table S1, while the structures of the drugs are shown in Table 1. Efficiency indices were also determined for all cases. Molecular surface area and polar surface area were not used because they can be sensitive to conformation.

The experimental free energies of binding were collected and the calculated free energies of binding were computed for each complex, using all the scoring functions as well as a selection of their components. Care was taken to use the experimental structure for calculating the docking score and only relaxations of this structure, or "docking in place" was performed, to maintain the same binding mode and pose for all programs (achieving complexes with electrostatic and van der Waals energies such as that in Fig. 1).

The results for several scoring functions and components for the 26 resulting protein-ligand complexes are shown in Table 2, and plotted results are shown in Figure 2a. Table 2 and Figure 2a show the difference in values for the experimental and calculated free energy of binding for each protein-drug complex. Chemscore and Goldscore are included in Figure 2a, even though they have positive scales (they return a positive value instead of a negative free energy). They were included as the negative of their value, i.e., −Goldscore (GS) and −Chemscore (CS), for the sake of comparison. XPc and SPc correspond to the Coulomb + van der Waals components of XP and SP, respectively. XP and SP correspond to XP and SP "refine" treatment, whereas SPi refers to SP "in place." ABE corresponds to autodock binding free energy, AIE to autodock intermolecular energy. DGe is the experimental calculated energy, DGb is a component of CS called ΔGbindGOLD. Some scoring functions have values that are closer to the experimental ones, and some follow the trend of the experimental values better.

The efficiency indices were then calculated for each scoring function value (as well as the selected components of the scoring functions) substituting the value for ΔG in ΔG/MW, ΔG/NHA, ΔG/NoC, log($-\Delta G/P$), and ΔG/W. The same efficiency indices were calculated for the experimentally determined ΔG. The means, medians, and $\pm$ standard deviations for all systems, as well as the complete tables are available in the Supporting Information Tables S2–S6. The plotted results for the molecular weight efficiency index for all scoring functions and experiment
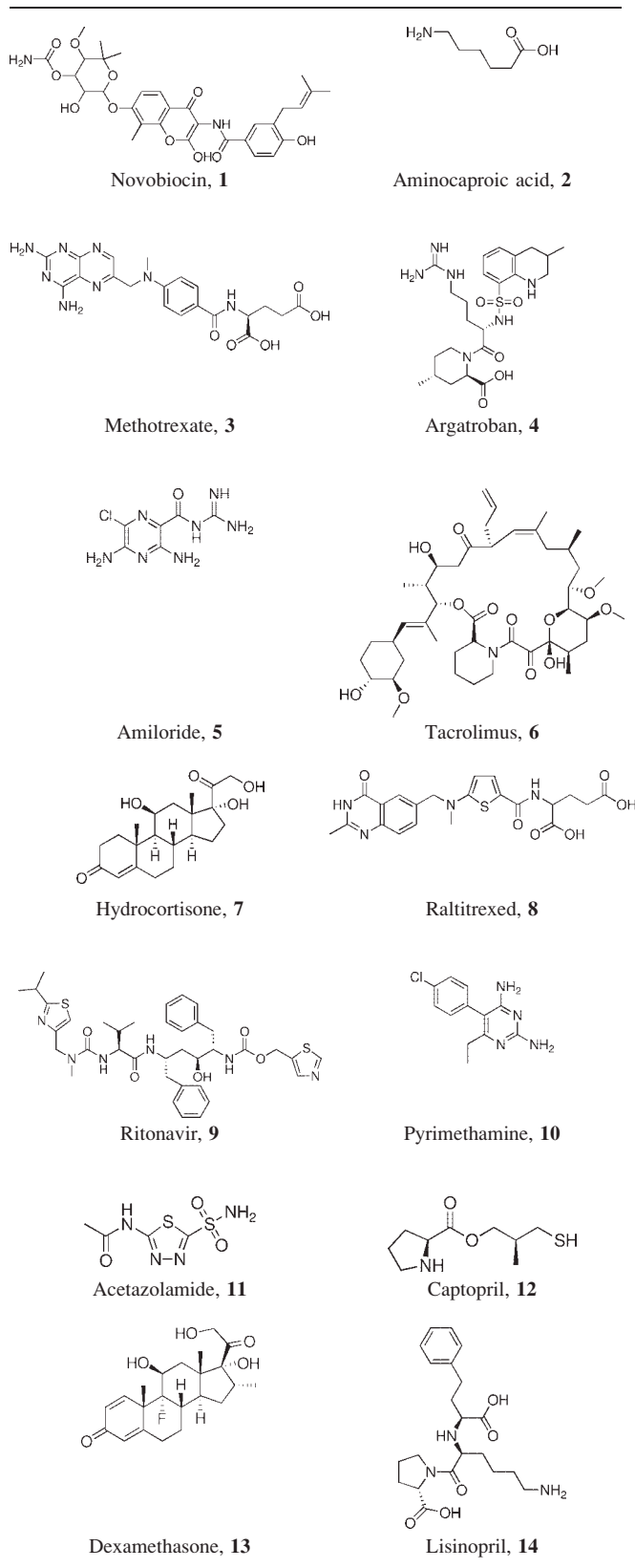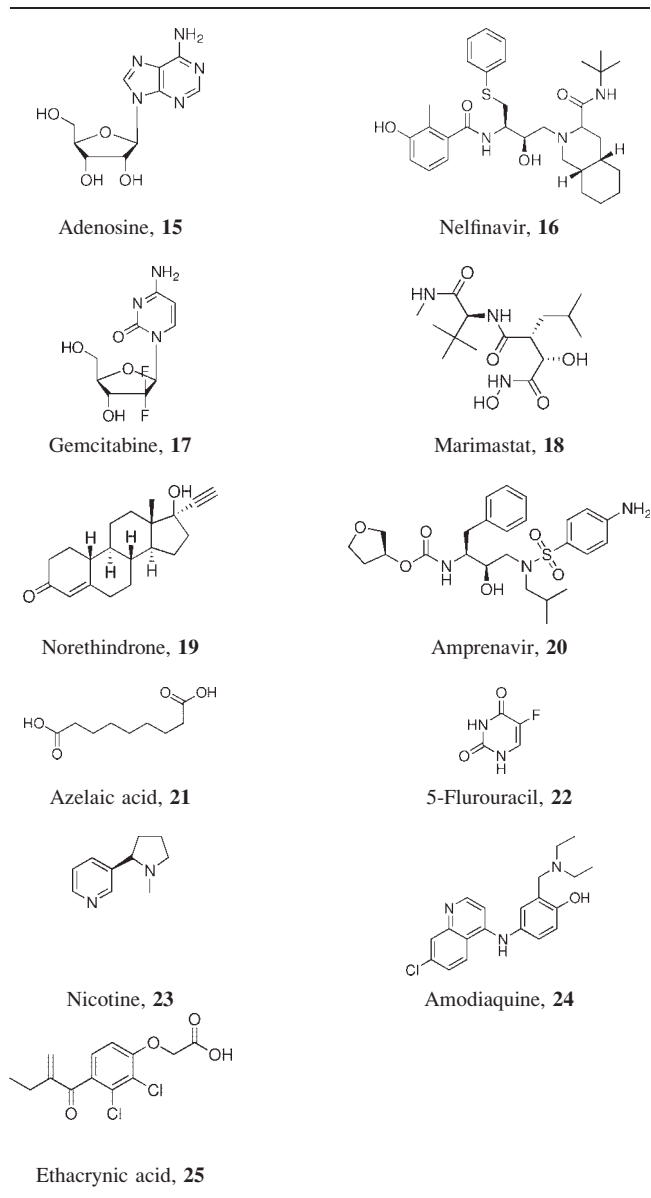
Table 1. Drug Structure Dataset.

Novobiocin, **1**

Aminocaproic acid, **2**

Methotrexate, **3**

Argatroban, **4**

Amiloride, **5**

Tacrolimus, **6**

Hydrocortisone, **7**

Raltitrexed, **8**

Ritonavir, **9**

Pyrimethamine, **10**

Acetazolamide, **11**

Captopril, **12**

Dexamethasone, **13**

Lisinopril, **14**

Table 1. (*Continued*).

Adenosine, **15**

Nelfinavir, **16**

Gemcitabine, **17**

Marimastat, **18**

Norethindrone, **19**

Amprenavir, **20**

Azelaic acid, **21**

5-Flurouracil, **22**

Nicotine, **23**

Amodiaquine, **24**

Ethacrynic acid, **25**

are shown in Figure 2b, while the Wiener index efficiency index is shown in Figure 2c.

As can be seen from Supporting Information Tables S1–S5 and Figure 2b (ΔG/MW), there is still some variation between the experimental efficiency indices and the calculated efficiency indices. However, Figure 2c (ΔG/W) shows how the experimental and calculated efficiency indices are now quite close. The linear regression correlation coefficients between the experimental and calculated binding energies, as well as between experimental and calculated efficiency indices were computed for all cases. They are shown in Table 3.
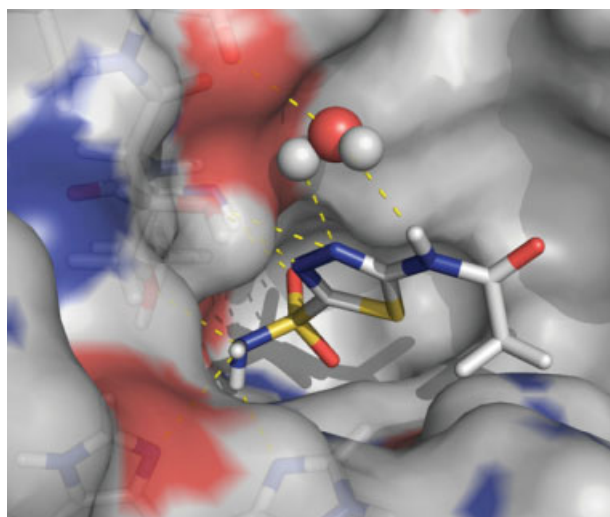
Figure 1. Complex of acetazolamide (11, sticks) with carbonic anhydrase X11 (surface and sticks) obtained by docking on crystal structure 1JD0. Nitrogen atoms in blue, oxygen in red, hydrogen in white, sulfur in yellow, hydrogen bonds as yellow dashes.

Some efficiency indices appear to be better than others for correlating experimental and calculated values. From Table 3, it can be seen that for some scoring functions, MW and NHA ei-

ther do not improve the results or provide only a modest improvement over the correlations with experimental values. The simple measure NoC (number of carbons) provides a good correlation for some of the scoring functions. This efficiency index is related to the nonpolar surface area, because the larger NoC a compound has, the larger its nonpolar surface is likely to be. Therefore, it may be providing an indirect measure of the desolvation energy for a molecule. The efficiency index $\log(-\Delta G/P)$ provides good correlations for all scoring functions between experimental and calculated values. This index is directly related to the permeability of a molecule. The efficiency index $\Delta G/W$ also improves all of the correlations. The p values in Table 3 show the probability that the corresponding F-statistic could have occurred by chance. All of them are below $\alpha = 0.05$, indicating that the regression models are useful in predicting the linear relationship with the experimental values (at a 95% confidence level). Efficiency indices, therefore, also appear to be able to introduce useful extra information in addition to the free energy of binding into a derived measurement.

As examples of the good linear correlations between experimental and calculated values, the plot of the experimental $\Delta G/$NoC versus calculated $\Delta G/$NoC for DGb (ΔGbindGOLD, a component of Chemscore) is shown in Figure 3a; experimental $\log(-\Delta G/P)$ versus calculated $\log(-\Delta G/P)$ for the same DGb is shown in Figure 3b; and the plot of the experimental $\Delta G/W$ versus calculated $\Delta G/W$ for DGb is shown in Figure 3c.

Table 2. Experimental and Calculated Free Energies of Binding (kcal mol^{-1}).[a]

PDB code	DGe	ABE	AIE	GS	CS	DGb	XP	SP	SPi
1aj6	−8.07	−9.97	−8.85	−44.84	−14.52	−17.88	−7.07	−7.69	−5.93
1cea	−6.76	−6.82	−6.82	−40.12	−20.72	−22.17	−8.18	−5.58	−6.86
1dhi	−9.90	−9.54	−9.54	−67.31	−22.29	−24.97	−9.46	−9.60	−9.00
1dhj	−8.93	−7.86	−10.56	−72.54	−25.83	−27.29	−8.72	−8.72	−7.60
1dwc	−10.10	−10.29	−10.29	−12.28	−27.79	−34.23	−11.08	−7.65	−6.07
1f5l	−7.19	−7.12	−7.52	−35.58	−17.86	−18.48	−6.92	−7.16	−6.57
1fkf	−12.81	−10.19	−12.04	−52.82	−35.20	−37.38	−11.33	−7.13	−6.26
1h61	−6.66	−6.81	−6.81	−24.87	−23.01	−23.99	−9.13	−5.90	−3.71
1hvy	−8.42	−8.33	−7.24	−46.14	−16.59	−17.50	−5.77	−6.87	−6.04
1hxw	−14.75	−15.12	−15.12	−90.22	−43.46	−47.69	−14.57	−12.24	−11.20
1j3j	−10.92	−7.50	−7.50	−52.20	−21.48	−25.36	−8.15	−6.55	−5.83
1jd0	−11.24	−5.87	−6.34	−41.68	−18.77	−22.77	−4.55	−4.22	−4.18
1m2x	−5.66	−14.66	−15.50	−53.91	−22.66	−24.30	−7.05	−9.82	−9.30
1m2z	−9.84	−10.24	−11.15	−46.61	−36.01	−38.04	−13.97	−9.44	−8.86
1o86	−13.04	−17.32	−21.14	−59.41	−34.82	−43.04	−12.90	−11.70	−10.69
1odi	−5.73	−5.34	−6.32	−48.17	−15.83	−16.93	−8.05	−7.90	−6.26
1ohr	−11.86	−12.57	−13.93	−52.87	−36.34	−39.09	−11.16	−9.43	−9.48
1p62	−6.35	−5.73	−5.73	−46.16	−18.62	−21.90	−12.56	−7.33	−5.76
1r55	−9.26	−9.37	−11.69	−52.77	−22.75	−35.60	−10.56	−9.47	−9.43
1sqn	−12.81	−10.07	−10.07	−60.74	−32.60	−35.49	−11.01	−8.51	−8.36
1t7j	−11.86	−10.34	−13.05	−72.31	−25.23	−29.21	−9.35	−7.68	−6.50
1tuf	−5.52	−7.11	−9.27	−23.67	−8.07	−11.63	−2.97	−4.36	−4.42
1upf	−6.27	−3.83	−3.83	−17.72	−10.25	−11.11	−5.65	−5.98	−5.30
1uw6	−10.01	−6.53	−6.78	−41.36	−25.63	−28.06	−4.97	−5.85	−5.94
2aou	−10.54	−10.35	−10.35	−24.79	−36.79	−39.81	−10.22	−9.00	−8.68
2gss	−6.73	−6.39	−7.63	−28.72	−18.02	−19.77	−6.45	−5.40	−5.57

[a]DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindGOLD; XP, XPrefine; SP, SPrefine; SPi, SP in place.
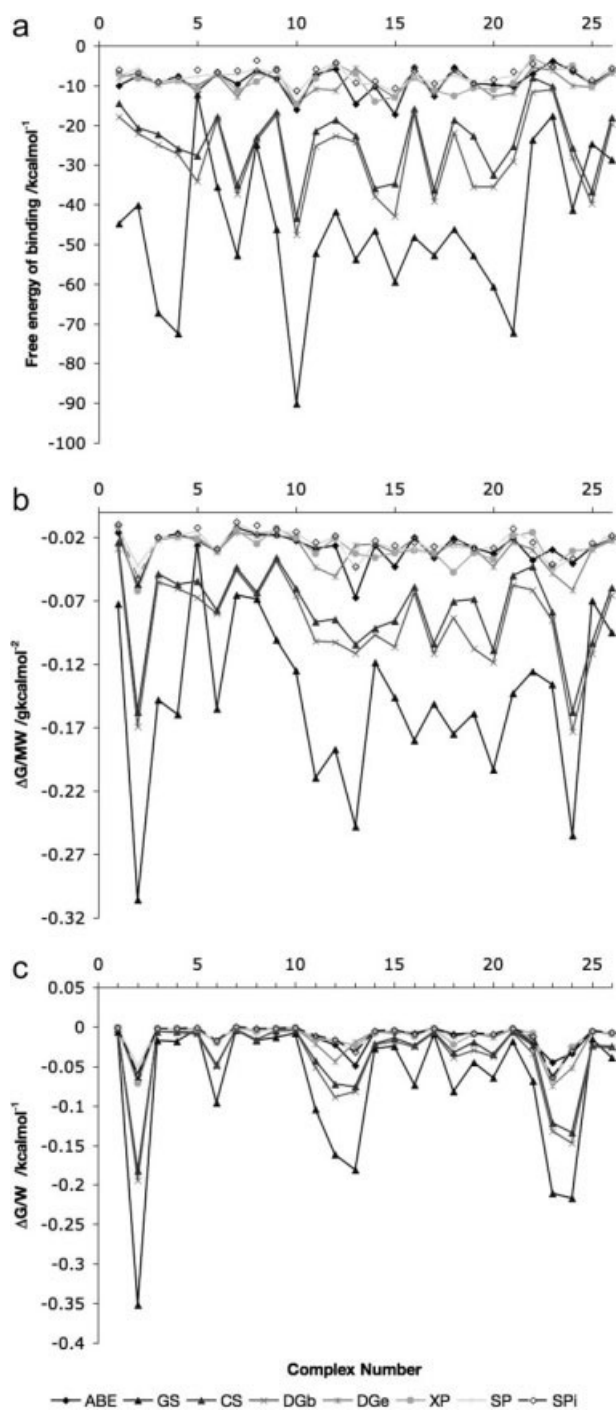
Figure 2. (a) Free energy of binding (ΔG) for each complex and several docking experiments, as well as determined by experiment. (b) Comparison of free energy of binding/molecular weight (ΔG/MW) efficiency indices for experiment and several scoring functions. (c) Comparison of free energy of binding/Wiener index (ΔG/W) efficiency indices for experiment and several scoring functions.

Since DGb is a component of CS, and therefore also of SP and XP, it is interesting to note that a component can have better correlation with experimental values than the full scoring

function. This can be due to the need of improvement in the extra features of the scoring function, such as the desolvation penalties and entropy corrections. As controls, the experimental free energy of binding was correlated against the MW, NHA, NoC, log P, and Wiener values. The results did not show any strong linear correlation, with the R^2 values being 0.323, 0.364, 0.404, 0.205, and 0.315, respectively. This means that the good correlations found between experimental and calculated efficiency indices are not spurious or redundant.

Linear regressions were also carried out between the simple DGe and all of the calculated efficiency indices, and they showed no linear correlation stronger than 0.1. Linear regressions were also calculated for all the efficiency indices against the molecular properties (MW, NHA, NoC, etc.) to test the dominance of these in the derived efficiency index, showing no strong linear correlation either, with most beneath 0.5, except SPi and SP (most SPi and SP correlation R^2s around 0.6, except SPi/W and SP/W vs. W, $R^2 = 0.2$). An exception for all the scoring functions was log P, which showed strong correlation between log($-\Delta G/P$) vs. log P of circa 0.99 in R^2. However, this effect was created by the logarithm function. If the simple $\Delta G/P$ was calculated instead, then all of the scoring functions had correlations between $\Delta G/P$ vs. P lower than 0.1. Indeed, this efficiency index is better suited than log($-\Delta G/P$), and also shows strong correlations between calculated and experimental efficiency indices as seen in Table 4.

The improvements in going from ΔG to the different efficiency indexes are shown in Figure 4, where it can be seen that some of the efficiency indices (ΔG/NoC, log($-\Delta G/P$), ΔG/W, and $\Delta G/P$) produce better improvements than others.

Figures 5a–5g show box plots for all of the distributions studied, and compares experimental and calculated efficiency indices where the spread between and within each series of data can be observed. The horizontal dark lines represent the median of the distributions, while the dark diamonds represent outliers. The plots of the free energy of binding have quite different spreads between the experimental and the calculated values, except for ABE, AIE, XP, SP, and SPi (Fig. 5a). ΔG/MW (Fig. 5b) and ΔG/NHA (Fig. 5c) indices do not change these spreads very much, while ΔG/NoC (Fig. 5d) already provides closer spreads between calculated and experimental values. Log($-\Delta G/P$) (Fig. 5e), ΔG/W (Fig. 5f), and $\Delta G/P$ (Fig. 5g) all show spreads that are now quite comparable between the experimental and calculated efficiency indices.

Shapiro normality tests were conducted for all the distributions studied, and they are shown in Supporting Information Table S8. Some of the distributions did not differ from a normal distribution with a 95% confidence limit and for these, Welch, independent, two-sided, t-tests were carried out between the experimental and calculated values (Table S9 in the Supporting Information). In the case of free energy of binding, for DGe, ABE, AIE, and XP, the test showed that the null hypothesis was true, i.e., that the true difference in means between the calculated and the experimental distributions is equal to zero, and they are comparable distributions. This was also the case for all the experimental and calculated log($-\Delta G/P$) efficiency indices.

For all of the distributions, Mann-Whitney U tests (a non-parametric test) were carried out to compare the experimental

Table 3. Linear Regression Correlation Coefficients and Statistics Between Experimental and Calculated Values for Binding Free Energy, $(y = ax + b)$ as well as Five Efficiency Indices.[a]

ScorF	ΔG	ΔG/MW	ΔG/NHA	ΔG/NoC	$\log(-\Delta G/P)$	ΔG/W
DGe	1	1	1	1	1	1
DGb	0.676, 50.2, $p < 0.001$	0.684, 51.9, $p < 0.001$	0.673, 49.4, $p < 0.001$	0.842, 127.7, $p < 0.001$	0.997, 9065, $p < 0.001$	0.885, 184.5, $p < 0.001$
CS	0.634, 41.6, $p < 0.001$	0.644, 43.4, $p < 0.001$	0.626, 40.3, $p < 0.001$	0.798, 94.7, $p < 0.001$	0.997, 7216, $p < 0.001$	0.870, 161.4, $p < 0.001$
GS	0.310, 10.8, 0.003	0.473, 21.6, $p < 0.001$	0.490, 23.0, $p < 0.001$	0.712, 59.4, $p < 0.001$	0.992, 2916, $p < 0.001$	0.822, 110.9, $p < 0.001$
XP	0.315, 11.0, 0.0029	0.319, 11.3, 0.0026	0.299, 10.2, 0.0038	0.450, 19.6, $p < 0.001$	0.994, 4341, $p < 0.001$	0.819, 108.6, $p < 0.001$
XPc	0.357, 13.3, 0.0012	0.362, 13.6, 0.0011	0.416, 17.1, $p < 0.001$	0.769, 79.8, $p < 0.001$	0.996, 5545, $p < 0.001$	0.877, 171.6, $p < 0.001$
SP	0.246, 7.82, 0.010	0.415, 17.0, $p < 0.001$	0.390, 15.3, $p < 0.001$	0.511, 25.1, $p < 0.001$	0.996, 6009.8, $p < 0.001$	0.856, 142.7, $p < 0.001$
SPc	0.387, 15.1, $p < 0.001$	0.363, 13.6, 0.0011	0.389, 15.3, $p < 0.001$	0.696, 55.1, $p < 0.001$	0.996, 5417, $p < 0.001$	0.912, 248.9, $p < 0.001$
SPi	0.261, 8.5, 0.0077	0.497, 23.7, $p < 0.001$	0.466, 20.9, $p < 0.001$	0.550, 29.3, $p < 0.001$	0.996, 5817, $p < 0.001$	0.864, 152.2, $p < 0.001$
ABE	0.347, 12.8, 0.0015	0.290, 9.8, 0.0046	0.273, 9.0, 0.0061	0.512, 25.2, $p < 0.001$	0.996, 5373, $p < 0.001$	0.743, 69.5, $p < 0.001$
AIE	0.318, 11.2, 0.0027	0.228, 7.1, 0.013	0.219, 6.7, 0.016	0.474, 21.6, $p < 0.001$	0.995, 4542, $p < 0.001$	0.718, 61.2, $p < 0.001$

R^2, F-statistic, and p values are given in the table.

[a]ScorF, scoring function; DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindGOLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.

and calculated distributions to assess whether two samples of observations come from the same distribution, including for all cases where the values were not normally distributed. The results are shown in Supporting Information Table S10. For free energy of binding, the test statistics W and p-values (p higher than 0.05, 95% confidence level) showed that there was no statistically significant difference between the experimental ΔG values and each of the calculated ABE, AIE, XP, and SP distributions. For ΔG/MW, there was no statistically significant difference between the experimental and each of the ABE, AIE, XP, SP, and SPi calculated distributions. This was also true for ΔG/NHA, ΔG/NoC, and ΔG/W index. For $\log(-\Delta G/P)$ and $\Delta G/P$, all of the calculated distributions had no statistically significant difference to the experimental one, for all of the scoring functions studied.

The equations between experimental and calculated values were then Y-scrambled with random numbers in the same range of values. Nearly all the R^2 values were markedly lower than for the unscrambled models (below 0.6). The only exception were the values of $\log(-\Delta G/P)$ which remained high even in the scrambled models. This indicates that this efficiency index is not particularly good for improving the correlations, since it cannot distinguish a true correlation from a random one, although the logarithm function was responsible for that behavior. Importantly, the efficiency index $\Delta G/P$ had low correlation values for the scrambled models, which indicates that it has reliability. From these scrambling results, we can see that there is a small component in the efficiency indices which improves the correlations with experimental values due to mathematical correction (that is, it is beneficial to have the values on the same scale), but it does not account for all of the improvement. This suggests that there may be physical underlying causes to the improvements, which depend on the normalizing measure incorporated into the efficiency index. The improvement effect may be due to description of the entropic part of the free energy of binding, through efficiency indices that describe the topology of a molecule (such as W).[32] Other efficiency indices such as NoC and P, may provide improvement through a description of the desolvation and of the permeability of a compound. For all scoring functions, the best efficiency indices effectively normalize the free energy derived indices, to give values closer to experiment.

Discussion

Efficiency indices can improve the outcome of docking scoring functions because they provide a closer agreement with experimental values. In addition, useful information related to the molecular properties of a molecule such as its lipophilicity P, or topology (described by W), can be incorporated into a single indicator. Some efficiency indices appeared to be better than others at improving the correlations. ΔG/NoC, ΔG/W, and $\Delta G/P$ are better than ΔG/MW or ΔG/NHA, and this effect was observed for all scoring functions.
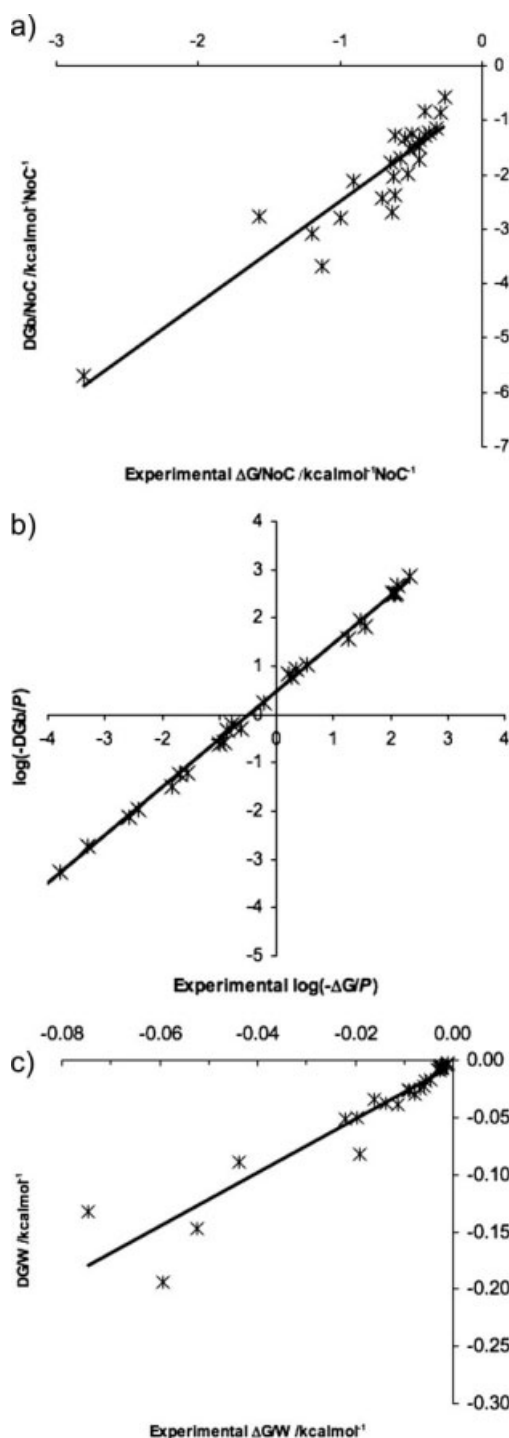
Figure 3. Experimental versus calculated values of the efficiency indices: (a) $\Delta G/NoC$ (free energy of binding/number of carbons) for DGb (ΔGbindGOLD) for 26 protein-drug complexes. $R^2 = 0.842$. (b) $\log(-\Delta G/P)$ (logarithm of $(-)$free energy of binding/octanol-water partition coefficient) for DGb for 26 protein-drug complexes. $R^2 = 0.997$. (c) $\Delta G/W$ (free energy of binding/Wiener index) for DGb for 26 protein-drug complexes. $R^2 = 0.885$.

To test the performance of efficiency indices with different types of compounds, the 25 ligands (in 26 protein-ligand complexes) were separated into two groups, small and large ligands if they were below or above the average MW and also by taking the 1st quartile (lowest 25%) and 3rd quartile (highest 25%). The same separation was conducted for polar and nonpolar ligands only now considering polar surface areas (PSA, in $Å^2$).[37] The sum of squares of the residuals (a measure of fitting error) were then recorded for each ligand complex for the differences between the calculated and the experimental value as: RSS $= \Sigma$(Experimental value $-$ Predicted value)$^2/n$, where n is the number of ligands, and the summation is over all the members in that group. The effect of molecular size on the efficiency indices were most marked for ABE, AIE, XP, and SP, where there was a large reduction of the difference between the errors for the small ligands compared to the large ones, using both separation methods. On average, small ligands had RSS errors of 10.19 kcal^2/mol^2 for binding free energy compared to 4.89 for large ligands. The efficiency indices markedly reduced this disparity till having equal differences between the errors for small and large ligands (average differences in RSS between small and large ligands: 0.0002 gkcal/mol^2 for ΔG/MW, 0.050 kcal/molNHA for ΔG/NHA, 0.234 kcal/molNoC for ΔG/NoC, 0.021 for $\log(-\Delta G/P)$, and 0.0001 kcal/mol for $\Delta G/W$). This applied to all efficiency indices except $\Delta G/P$ where only SP/P produced the smallest differences (0.91). Nonpolar ligands (i.e., those with a small polar surface area) were also at a disadvantage compared to polar ones (large polar surface area). Using the 1st and 3rd quartiles, the efficiency indices (except $\Delta G/P$) for ABE, AIE, XP, and SP corrected this bias by reducing the differences in errors from averages of 9.59 kcal^2/mol^2 in binding free energy for nonpolar ligands and 4.27 for polar ligands, so that the differences between

Table 4. Linear Regression Correlation Coefficients and Statistics Between Experimental and Calculated Values for Binding Free Energy/Octanol-Water Partition Coefficient ($\Delta G/P$) Efficiency Index.[a]

Scoring function or component	$\Delta G/P$	Scoring function or component	$\Delta G/P$
DGe	1	DGe	1
DGb	0.981, 1247.9, $3.31e^{-22}$	SPc	0.936, 350.3, $8.9e^{-16}$
CS	0.989, 2206.8, $3.9e^{-25}$	SP	0.986, 1686.1, $9.47e^{-24}$
GS	0.926, 299.7, $4.61e^{-15}$	SPi	0.995, 4547, $7.09e^{-29}$
XP	0.879, 174.9, $1.62e^{-12}$	ABE	0.960, 578.1, $2.6e^{-18}$
XPc	0.924, 292.8, $6.0e^{-14}$	AIE	0.930, 317.2, $2.4e^{-17}$

R^2, F-statistic, and p values are given in the table.
[a]DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, $-$Goldscore; CS, $-$Chemscore; DGb, ΔG_{bind}GOLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.
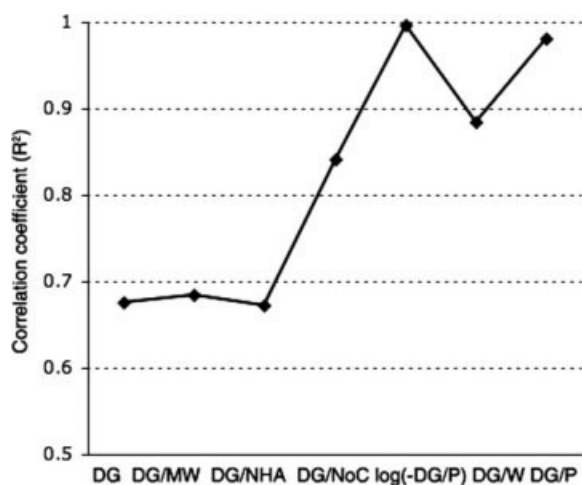
Figure 4. Correlation between experimental and calculated efficiency indices for ΔG, $\Delta G/MW$, $\Delta G/NHA$, $\Delta G/NoC$, $\log(-\Delta G/P)$, $\Delta G/W$, and $\Delta G/P$ for the scoring function component DGb (ΔGbindGOLD) for 26 protein-drug complexes.

the errors between experimental and predicted values were small and similar for both classes of ligands (average differences in RSS between nonpolar and polar ligands: 0.0003 gkcal/mol^2 for $\Delta G/MW$, 0.027 kcal/molNHA for $\Delta G/NHA$, 0.064 kcal/molNoC for $\Delta G/NoC$, 0.009 for $\log(-\Delta G/P)$, and 0.0001 kcal/mol for $\Delta G/W$). Using below and above average PSA to divide the groups, only ABE, AIE and XP showed this effect. The original bias may have risen due to possible overestimation of ligand-protein hydrogen bonding interactions by the scoring functions, or due to inadequate desolvation energy calculation of the nonpolar ligands by the scoring functions.

The drugs shown in Table 1 include large and small size ligands. It is typical for scoring functions to overestimate the binding energy of a compound because they are additive in nature: the larger the ligand, the more protein-ligand interactions it will have. However, the introduction of normalizing ligand efficiency measures allow for the smaller size ligands to be compared positively with larger size ligands because the binding energy is divided by a value which can be related to the molecular size. In a docking or virtual screening experiment, molecules with a large number of carbons are no longer favored. In this way, efficiency indices can repair the errors introduced by the bias of scoring functions toward large size ligands due to errors in the calculation of entropy and desolvation energies. It is promising that the effect was seen on all the scoring functions and programs.

Molecules with extremely high or low values of hydrophobicity or hydrophilicity, that is, with extreme values of P or $\log P$, can be removed through filters before docking. The $\Delta G/P$ efficiency index will also penalize those with borderline values, in addition to having low calculated free energy of binding. Thus, molecules with unfavorable permeability values can be readily detected, in combination with the binding free energy. If ranges for values are established for the different efficiency indices (see for example Hetényi et al.,[33] and also in this present work),

these can tell whether a compound's calculated efficiency index is in a favorable range. New efficiency indices can be compared and tested in a manner analogous to the present work.

Conclusions

We have shown that simple ligand efficiency indices can aid the drug design process by providing better comparisons of calculated and experimental values of binding energy. This may increase the accuracy and reliability of docking programs. We observed that efficiency indices also add information to the binding free energy into a single indicator. Permeability of a compound, for example, can be assessed at the same time as the binding affinity in an efficiency index such as $\Delta G/P$, especially if filters are applied to remove compounds with extreme values. Entropy of a compound may be assessed by $\Delta G/W$ values. $\Delta G/NoC$, although simple, is also an effective efficiency index for improving the trend between experimental and calculated values. $\Delta G/P$ and $\Delta G/W$ produced the best results, together with $\Delta G/NoC$. These efficiency indices can be applied across different docking programs, scoring functions, or even components of these. They can also be calculated quickly, likely on the fly. Compounds can be ranked based on efficiency indices that may include data such as absorption and metabolic properties in addition to the free energy of binding, and in this way allow for the selection of molecules that satisfy several criteria in parallel.

Computational Methods

The structures of protein-drug complexes and their experimental inhibition constants (Ki) were collected from the PDBbind database v2005,[38,39] which contains protein-ligand complex structural data from the Protein Data Bank (PDB)[40] as well as experimental Ki determined for those systems. The collection of all small-molecule approved drugs was obtained from the DrugBank database,[41] which contains data on drugs approved by the FDA (U.S. Food and Drug Administration agency). Programs written in Python were used to extract the ligand names (HET-ID) from the PDBbind database and to query them in the DrugBank collection to identify those ligands that are approved drugs. All results were verified visually. The experimental ΔG was computed with $\Delta G = -RT\ln K$, using $T = 25°C$ (298.15 K), and $R = 1.987$ cal/Kmol. The program XLOGP v2.0[42] was used for calculating the octanol/water partition coefficient ($\log P$) by an atom-additive method including correction factors.

Docking programs differ by the scoring functions they contain, as well as the way of minimizing the function values. In our present study, we focused on three main programs that are widely available and used by computational and medicinal chemists: GOLD v.3.1,[43] Glide v.4.5,[44] and Autodock4.[45] Their scoring functions and docking methods are shown in Supporting Information Table S11. GOLD v.3.1[43] uses a genetic algorithm to find the best ligand positioning in a binding site. It can use two scoring functions: Goldscore[43] and Chemscore.[46] Chemscore has a component called ΔGbinding (DGb), which was also used for our correlations. Parameters for runs were: run_flag =
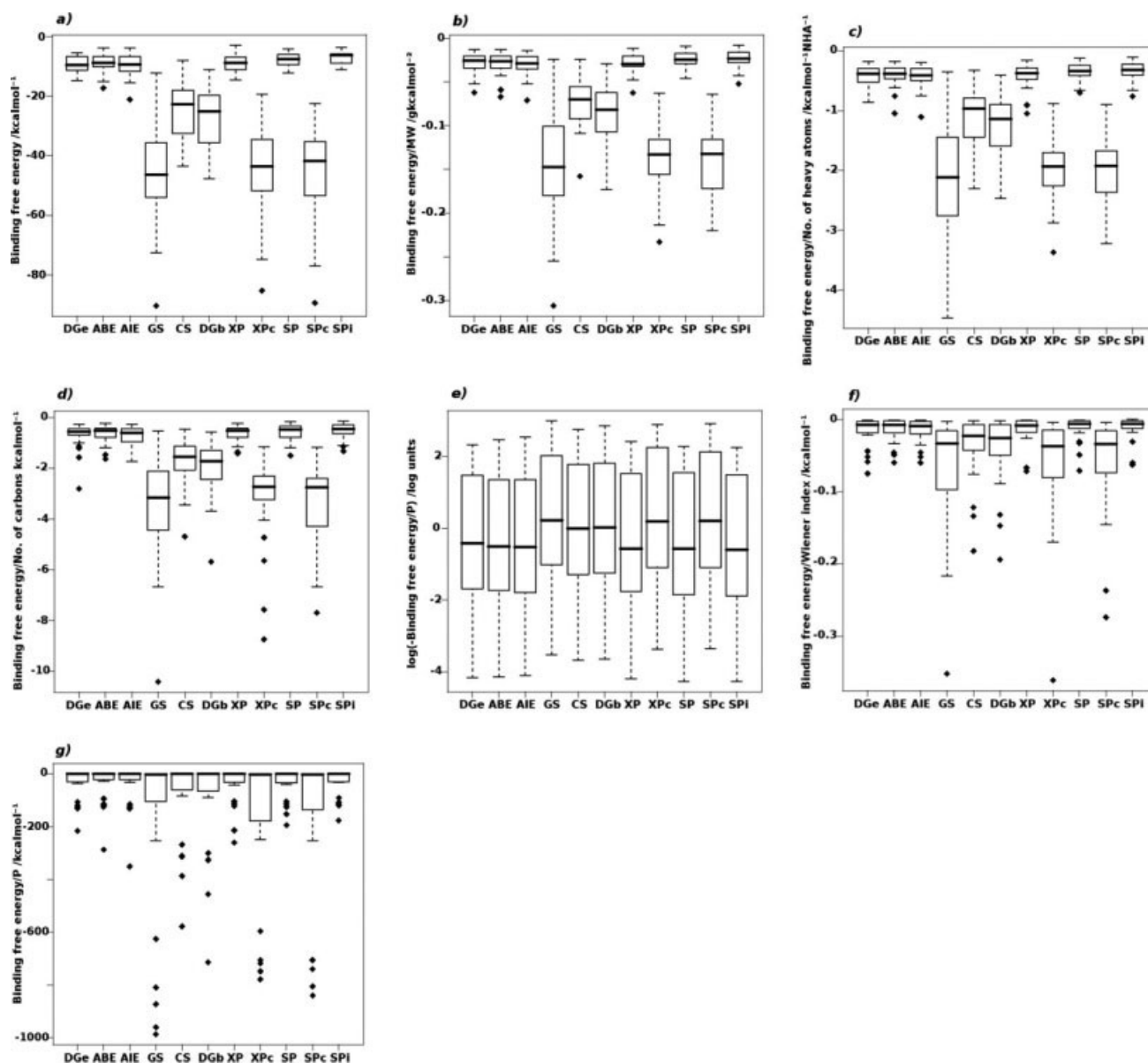
Figure 5. Box plot comparisons of free energies and efficiency indices for experiment and several scoring functions: (a) Free energy of binding (ΔG). (b) Free energy of binding/molecular weight (ΔG/MW). (c) Free energy of binding/number of heavy atoms (ΔG/NHA) efficiency index. (d) Free energy of binding/number of carbons (ΔG/NoC) efficiency index. (e) Logarithm of the (changed sign) free energy of binding/octanol-water partition coefficient ($\log(-\Delta G/P)$) efficiency index. (f) Free energy of binding/Wiener index ($\Delta G/W$) efficiency index. (g) Free energy of binding/octanol-water partition coefficient ($\Delta G/P$) efficiency index. DGe, experimental binding free energy; ABE, autodock binding free energy; AIE, autodock intermolecular energy; GS, −Goldscore; CS, −Chemscore; DGb, ΔGbindG-OLD; XP, XPrefine; XPc, XP_CvdW (Coulomb and van der Waals components of XP); SP, SPrefine; SPc, SP_CvdW (Coulomb and van der Waals components of SP); SPi, SP in place.

RESCORE, in addition to default parameters for the genetic algorithm. Waters were switched to ON.

Glide v4.5 (2007) uses a hierarchical search, and has the scoring functions XP and SP,[44] which are a proprietary modification of Chemscore.[46] In addition, we also employed the com-

ponent C_vdW (a combination of Coulomb and van der Waals terms). Default parameters for runs were used.

Autodock v4.0 also uses a genetic algorithm to find for the best solutions for docked ligands. It uses one scoring function, which produces a binding free energy (ABE).[47] We also

employed the component of intermolecular energy (AIE). Parameters used that were different than default values: spacing = 0.375 Å, npts = 40 40 40, ga_pop_size = 150, ga_num_evals = 20,000,000, ga_num_generations = 27,000, tran0 coordinate equal to the "about coordinates", quat0 = 1. 0. 0. 0., and dihe = 0.

Protein and ligand structures already contained hydrogens from the PDBBind dataset. Protein structures were used including the metal atoms and select water molecules that were interacting with protein and ligand in the binding site. The "toggle" setting was used for these special bridge water molecules in GOLD. Docking runs were calculated both including and excluding select crystallographic water molecules. The case which produced a binding energy closest to the experimental was kept. Most of the complexes which included select water molecules had a small effect on the binding energy and efficiency indices as they differed by less than 1 kcal/mol from the "dry" cases in binding free energy, as well as being in the same range and evenly distributed for binding free energy and efficiency indices as the cases without water molecules. The complete list of water molecules is shown in Supporting Information Table S7. Water molecules were included only if they had medium to low crystallographic B-factors, made contacts with the protein, were within 4.5 Å of the ligand, and were at least partially occluded from bulk solvent since these tightly bound water molecules have a higher chance of remaining bound to the protein (remaining conserved in several protein structures),[13–15] and can be considered an integral part of the protein-ligand complex. As such, these specially selected crystallographic water molecules are included in the binding free energy, as well as in the efficiency indices. There was no additional water inclusion or removal when calculating the efficiency indices, which take their binding energy direct from the complex. Efficiency indices are unique to each protein and each ligand in a biomolecular complex, although general trends and ranges can be observed across complexes. Methotrexate (**3**), for example, forms two complexes with different proteins in the dataset, consequently with different binding free energies and efficiency indices. Most of the protein-drug structures which included bridge water molecules mediating their interaction had high crystal structure resolutions, from 1.4 Å and on average lower than 2 Å (median of 1.95 Å) which may increase the probability of detecting reliable water molecule electron density.[48] They included a wide diversity of ligands, though the exposed, shallow complex of the small, relatively nonpolar aminocaproic ligand (**2**) did not have bridging waters, nor did the completely buried dexamethasone (**13**).

Complexes were prepared for the dockings by minimizing in water with generalized Born (GB) implicit solvation and a steepest descent method to a gradient threshold of 239 kcal/molnm, followed by a minimization in water (GB) with a truncated-Newton conjugated gradient method to a gradient threshold of 143.4 kcal/molnm using MacroModel.[49] Statistical tests and box plots were performed using the package R for statistical computing.[50] Marvin Calculator Plug-ins were used for the calculation of ligand molecular formulas and molecular mass (MW).[37]

References

1. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Nat Rev Drug Discov 2004, 3, 935.
2. Coupez, B.; Lewis, R. A. Curr Med Chem 2006, 13, 2995.
3. Klebe, G. Drug Discov Today 2006, 11, 580.
4. Huang, N.; Jacobson, M. P. Curr Opin Drug Discov Dev 2007, 10, 325.
5. Waszkowycz, B. Drug Discov Today 2008, 13, 219.
6. Pierce, A. C.; Jacobs, M.; Stuver-Moody, C. J Med Chem 2008, 51, 1972.
7. Vigers, G. P. A.; Rizzi, J. P. J Med Chem 2004, 47, 80.
8. Chang, C. E. A.; Chen, W.; Gilson, M. K. Proc Natl Acad Sci USA 2007, 104, 1534.
9. Velec, H. F. G.; Gohlke, H.; Klebe, G. J Med Chem 2005, 48, 6296.
10. Pfeffer, P.; Gohlke, H. J Chem Inf Model 2007, 47, 1868.
11. Raub, S.; Steffen, A.; Klamper, A.; Marian, C. M. J Chem Inf Model 2008, 48, 1492.
12. Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. Proteins 2008, 73, 395.
13. García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. J Mol Model 2003, 9, 172.
14. Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. J Mol Biol 2003, 326, 621.
15. García-Sosa, A. T.; Firth-Clark, S.; Mancera, R. L. J Chem Inf Model 2005, 45, 624.
16. Li, Z.; Lazaridis, T. J Phys Chem B 2006, 110, 1464.
17. Li, Z.; Lazaridis, T. Phys Chem Chem Phys 2007, 9, 573.
18. Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Med Res Rev 2006, 26, 531.
19. Amaro, R. E.; Baron, R.; McCammon, J. A. J Comput Aided Mol Des 2008, 22, 693.
20. Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Constantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. J Med Chem 2008, 51, 6237.
21. Chen, W.; Chang, C. E.; Gilson, M. K. Biophys J 2004, 87, 3035.
22. Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. J Chem Inf Model 2005, 45, 1134.
23. Wang, J. M.; Kang, X. S.; Kuntz, I. D.; Kollman, P. A. J Med Chem 2005, 48, 2432.
24. Lyne, P. D.; Lamb, M. L.; Saeh, J. C. J Med Chem 2006, 49, 4805.
25. Guimarães, C. R. W.; Cardozo, M. J Chem Inf Model 2008, 48, 958.
26. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Proteins 2005, 60, 325.
27. Warren, G. L.; Andrew, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lamber, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. J Med Chem 2006, 49, 5912.
28. Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. Proc Natl Acad Sci USA 1999, 96, 9997.
29. Hopkins, A. L.; Groom, C. R.; Alex, A. Drug Discov Today 2004, 9, 430.
30. Abad-Zapatero, C.; Metz, J. T. Drug Discov Today 2005, 10, 464.
31. Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. J Med Chem 2008, 51, 2432.
32. Hetényi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. J Am Chem Soc 2006, 128, 1233.
33. Hetényi, C.; Maran, U.; García-Sosa, A. T.; Karelson, M. Bioinformatics 2007, 23, 2678.
34. García-Sosa, A. T.; Sild, S.; Maran, U. J Chem Inf Model 2008, 40, 2074.
35. Wells, J. A.; McClendon, C. L. Nature 2007, 450, 1001.

36. Leeson, P. D.; Springthorpe, B. Nat Rev Drug Discov 2007, 6, 881.

37. Marvin v4.8.1. 2007. ChemAxon. Available at: http://www.chemaxon.com (accessed on October 30, 2008).

38. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. J Med Chem 2004, 47, 2977.

39. Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C.-Y.; Wang, S. M. J Med Chem 2005, 48, 4111.

40. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.

41. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Nucleic Acids Res 2006, 34, D668.

42. Wang, R. X.; Gao, Y.; Lai, L. H. Perspect Drug Discov 2000, 19, 47.

43. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Mol Biol 1997, 267, 727.

44. Schrödinger, LLC. Glide Version 4.5; Schrödinger, LLC: New York, 2007.

45. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. J Comput Chem 1998, 19, 1639.

46. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. J Comput Aided Mol Des 1997, 11, 425.

47. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. J Comput Chem 2007, 28, 1145.

48. Davis, A. M.; Teague, S. J.; Kleywegt, G. J. Angew Chem Int Ed Engl 2003, 42, 2718.

49. Schrödinger, LLC. Macromodel v9.5; Schrödinger, LLC: New York, 2007.

50. The R Project for Statistical Computing. Available at: http://www.r-project.org (accessed on October 30, 2008).