JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Robust Scoring Functions for Protein−Ligand Interactions with Quantum Chemical Charge Models

Jui-Chih Wang,[†] Jung-Hsin Lin,[*,§,‡,∥] Chung-Ming Chen,[†] Alex L. Perryman,[⊥] and Arthur J. Olson[⊥]

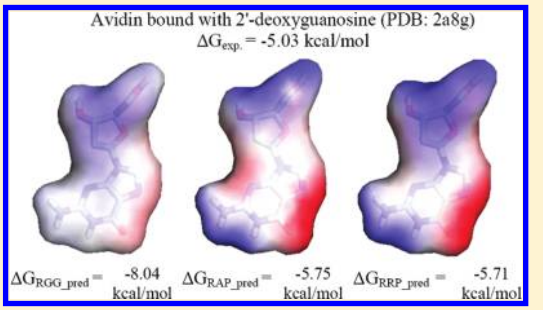
[†]Institute of Biomedical Engineering and [‡]School of Pharmacy, National Taiwan University, Taipei, Taiwan

[§]Division of Mechanics, Research Center for Applied Sciences and [∥]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

[⊥]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California, United States

Supporting Information

ABSTRACT: Ordinary least-squares (OLS) regression has been used widely for constructing the scoring functions for protein−ligand interactions. However, OLS is very sensitive to the existence of outliers, and models constructed using it are easily affected by the outliers or even the choice of the data set. On the other hand, determination of atomic charges is regarded as of central importance, because the electrostatic interaction is known to be a key contributing factor for biomolecular association. In the development of the AutoDock4 scoring function, only OLS was conducted, and the simple Gasteiger method was adopted. It is therefore of considerable interest to see whether more rigorous charge models could improve the statistical performance of the AutoDock4 scoring function. In this study, we have employed two well-established quantum chemical approaches, namely the restrained electrostatic potential (RESP) and the Austin-model 1-bond charge correction (AM1-BCC) methods, to obtain atomic partial charges, and we have compared how different charge models affect the performance of AutoDock4 scoring functions. In combination with robust regression analysis and outlier exclusion, our new protein−ligand free energy regression model with AM1-BCC charges for ligands and Amber99SB charges for proteins achieve lowest root-mean-squared error of 1.637 kcal/mol for the training set of 147 complexes and 2.176 kcal/mol for the external test set of 1427 complexes. The assessment for binding pose prediction with the 100 external decoy sets indicates very high success rate of 87% with the criteria of predicted root-mean-squared deviation of less than 2 Å. The success rates and statistical performance of our robust scoring functions are only weakly class-dependent (hydrophobic, hydrophilic, or mixed).



INTRODUCTION

Evaluation of the binding affinities of drug-like molecules with the target proteins is crucial for discriminating drug candidates from weak-binding or even nonbinding small molecules. Most, if not all, computational docking methods rely greatly on empirical or semiempirical scoring functions to evaluate protein−ligand interactions. Rigorous statistical mechanical approaches for evaluation of binding free energies are theoretically most satisfactory,[1,2] but such approaches are computationally too demanding for virtual screening. The simplest forms of evaluating protein−ligand binding affinity are empirical scoring functions[3−6] based on the quantitative structure−activity relationships (QSAR) approach pioneered by Hansch[7] or the semiempirical models with molecular mechanics-based energetics.[8−11] Common in these approaches is multivariate regression. Semiempirical models based on molecular mechanics have the advantages of easier rational interpretation of binding modes, and they are more sensitive to protein conformational changes. This is particularly important when protein dynamics and flexibility are to be accommodated.[12−14] Frequently used energetic terms include dispersion/repulsion (i.e., van der Waals energy), electrostatic energy, hydrogen-bond energy, desolvation energy, hydrophobic interaction, torsional entropy, etc.[8,9] Among these terms, the atomic partial charges of biomolecules are considered of central importance, because they are essential for evaluation of the long-ranged electrostatic interaction, which is known to be a key contributing factor for biomolecular association. Due to the extremely low computational cost, especially when facilitated with precalculated grid maps, current molecular docking programs often use regression models with distance-dependent molecular descriptors or energy terms to predict the possible binding poses of a small molecule, to evaluate its binding affinity, or to use for large-scale virtual chemical library screening for rapidly limiting the chemical space and for subsequent identification of potential drugs.

Intuitively, inclusion of more energetic terms or molecular descriptors in a scoring function may provide a more complete description of protein−ligand interactions and a more accurate binding free energy model. However, the introduction of many variables in a regression model can often lead to the overfitting problem,[15] which is caused by the vast emptiness of

Received: May 18, 2011
Published: September 21, 2011

high-dimensional multivariate space. On the other hand, the selection of molecular descriptors or energetic terms will also dictate the performance and applicability of such free energy models.[16]

The AutoDock4 scoring function[9] is a semiempirical scoring function that is embedded in the automated molecular docking software package, AutoDock, and has been widely adopted in virtual screening of drug candidates and prediction of ligand binding poses in protein pockets. The energetic terms in the AutoDock4 scoring function include the van der Waals interaction, the electrostatic interaction, the hydrogen-bonding interaction, the desolvation free energy, and the loss of ligand torsional entropy upon binding. The atomic charges used to evaluate the electrostatics energy term of the AutoDock4 were prepared using the Gasteiger charge model,[17] whose primary advantages lie in its simplicity and speed. However, such charge calculations can generate atomic charges that are less accurate than those determined by quantum chemical methods. For example, the dipole moment of the well-known polar molecule dimethyl sulfoxide (DMSO) calculated by the Gasteiger model is only 2.96 Debye (D), which is quite different from the results of restrained electrostatic potential (RESP) (4.61 D) and Austin-model 1-bond charge correction (AM1-BCC) (4.57 D). (The dipole moment of DMSO in solution can be estimated[18] from its measured dipole moment in vacuum (3.96 D)[19] to be about 4.7 D.) Due to the increase of computing power, ab initio quantum mechanical calculations can now be performed routinely, and some recent studies have indicated that docking calculations using more accurate atomic charges can indeed predict binding poses more accurately.[20−22] However, in these studies the more advanced charge models were not employed to construct new scoring functions, which may weaken the assertion that the more advanced charge models have superior predictive power. If a charge model is qualitatively and quantitatively different from the charge model used to develop a scoring function, one can expect large prediction errors. In other words, simply employing more accurate atomic charge models should not generally lead to better predictions of binding poses and binding affinities. Weighting coefficients of scoring functions in the empirical QSAR models need to be recalibrated.

In this study, we report our investigation of the influence of different charge models on the AutoDock4 scoring function, in order to see how different charge models affect the performance of the same functional form of the AutoDock4 scoring function. Our ordinary least-squares regression analyses indicated that AM1-BCC or RESP charges for ligands in combination with Amber99SB charges for proteins yielded lower root-mean-squared errors (RMSE). Because proper outlier exclusion is also important for calibration of empirical scoring functions, we have performed robust regressions and analyzed outliers. The performances of robust regression in several QSAR modeling tasks have been reported,[23] and it was concluded that robust regression is always better than ordinary least-squares (OLS) regression. Recently, we performed robust regression to delineate the influence of the data set on the calibration of empirical scoring functions for protein−ligand interactions.[24] Here we compare the statistical performance of the new robust regression models with different charge combinations on both the training set and a large external test set. We have also tested the performance of binding pose prediction of the new robust models on the 100 external decoy sets[25] that have been widely used in the assessment of protein−ligand scoring functions[26,27] as well as on a new decoy set of 195 complexes.[28]

METHODS

Functional Form of AutoDock4 Scoring Function. Improvements of the AutoDock4 (AD4) scoring function[9] over its predecessor, AutoDock3,[8] include a refined functional form of the desolvation energy, more atom types, and a significantly larger training data set. The AD4 scoring function comprises five energetic terms: the van der Waals interaction, the hydrogen bonding interaction, the electrostatic interaction, the desolvation energy, and the torsional entropy. The AD4 scoring function predicts the binding free energy with the following formula:

$$\Delta G_{\text{bind}} = W_{\text{vdW}} \times \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{\text{H-bond}} \times \sum_{i,j} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{\text{estat}} \times \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{\text{desol}} \times \sum_{i,j} (S_i V_j + S_j V_i)\, e^{(-r_{ij}^2/2\sigma^2)} + W_{\text{tor}} \times N_{\text{tors}}$$

The weighting coefficients W_i were obtained by regression analysis of the experimental binding affinity information collected in Ligand Protein Database (LPDB).[29] The van der Waals potential energy is a typical 12-6 form, where parameter A_{ij} and B_{ij} were adopted from the '84 Amber force field.[30] The hydrogen-bonding term is based on a 12-10 potential, weighted by a directional term, $E(t)$. The electrostatic interaction is calculated with a screened Coulomb potential.[31] The desolvation term is included by calculating the surrounding volume of an atom (V_i), weighted by the atomic solvation parameter (S_i) and an exponential term with a distance weighting factor σ (0.35 Å in AutoDock4). The final term represents the torsional entropy term, which is calculated simply by counting the number of rotatable bonds of a ligand.[32]

Charge Models. In this work, we focus on the two charge models, RESP[33] and AM1-BCC,[34] that have been used widely in molecular dynamics simulations with the AMBER force field. RESP[33] is a two-stage restrained electrostatic fit charge model. While the geometry of the molecule was taken from experimental structures, the quantum mechanical electrostatic potentials (ESP) based on the 6-31G* basis set were evaluated at the shells of points with the density of one point/Å^2 at each of 1.4, 1.6, 1.8, and 2.0 times the van der Waals radii of the molecule. Then, atom-centered model charges were derived by minimizing the differences between the reproduced ESP and the original quantum mechanical (QM) ESP plus the deviation from the minimum of a hyperbolic restraint function. In the first stage of the fitting process, no forced symmetry is applied, and a weak restraint is used. In the second stage, the charges on equivalent atoms are forced to be the same, and a strong restraint is used. QM calculations were performed by GAUSSIAN 09[35] at the Hartree−Fock (HF) level with the 6-31G* basis set.[36] The RESP atomic charges were computed by using Antechamber of the AMBER 11 suite based on the GAUSSIAN output file and were saved as the Tripos Mol2 format. Subsequently, the ADT program (prepare_ligand4.py) enables the conversion from the mol2 file to the pdbqt file format with the RESP atomic charges obtained.

As a semiempirical approach, AM1-BCC is a quick and efficient atomic charge model that aims to achieve the accuracy of RESP.[34,37] The AM1 charges were first calculated from the MOPAC 6 program for the individual molecule. The "am1bcc" program of the Antechamber package assigned bond and atom types and then performed bond charge corrections (BCCs) that were parametrized against the HF/6-31G* electrostatic potentials of a set of training compounds. The Tripos mol2 file with AM1-BCC atomic charges was saved by the Antechamber program and then converted to the ligand pdbqt file by the ADT program (prepare_ligand4.py).

The atomic charges of proteins were retrieved from the AMBER parm99SB force field parameters, which were mainly derived by the RESP methodology.[18,38,39] The residue name was assigned to comply with the Amber naming scheme, e.g., histidine with hydrogens on both nitrogens (HIP), histidine with hydrogen on the ε nitrogen (HIE), histidine with hydrogen on the δ nitrogen (HID), disulfide bonded cysteine (CYX), and so on. Subsequently, the LEaP program of AMBER 11 was employed to produce the coordinates and parameter/topology files, which were used to generate the "pqr" files with atomic charges and radii with the "ambpdb" program. These atomic charges then substituted the charges in the original "pdbqt" files that were generated by the ADT program (prepare_receptor4.py) with default settings.

Preparation of Protein and Ligand Structural Files. Before the calculations of atomic charges for the ligands, first, hydrogen atoms need to be added, and the net charges of the molecules should be determined. The ligand structural information was extracted from a complex with the form of biological assembly in Protein Data Bank, and then hydrogen atoms were added by OpenBabel,[40] net charges calculated by the "estimateFormalCharge" function in Chimera.[41] Because OpenBabel does not always assign correct protonation states, we further checked the protonation assignment of each ligand carefully and corrected its mistakes by our in-house scripts.

The protonated states of receptors and ligands were obtained from a previous preparation of Huey et al.[9] Similarly, ligands were optimized by using the local search capability of AutoDock[8] to avoid too close contact in the crystallographic atomic structural model.

It should be noted that many cofactors exist in several complexes of LPDB. These cofactors are neither amino acids nor parts of ligands, but they are often required for biological activity. Occasionally, these cofactors are located near ligands and are also inside the grid box of precalculated protein−ligand interactions. In the original version of AutoDock4 scoring function, the atomic charges of cofactors were also determined by the Gasteiger method.[9] To be consistent in the charge models of ligands in this work, the RESP and AM1-BCC models were also utilized on these cofactors. Some of the RESP charges of the cofactors can be retrieved from the literature: the charges of heme group were obtained from Autenrieth et al.[42] (for cytochrome c) and Oda et al.[43] (for cytochrome P450).

Because the AutoDock4 scoring function was calibrated for united atom models, the nonpolar hydrogen atoms were merged, and the united atom charges calculated by the ADT program (prepare_ligand4.py or prepare_receptor4.py). In the next sections of this article, the abbreviations "AP" for AM1-BCC (ligand)/Amber PARM99SB (protein) and "RP" for RESP (ligand)/Amber PARM99SB (protein) have been adopted.

Calculation of Energetic Terms. To evaluate the various energetic terms, the grid maps of different atom types of ligand were constructed by the "autogrid4" programs, with a grid spacing of 0.375 Å. The grid center was positioned at the geometrical center of a ligand, and the grid box was adjusted according to the size of a ligand, plus 22.5 Å. Subsequently, the "autodock4" program was used to calculate the energetic terms of protein−ligand interactions by setting the parameter as "epdb" in the AutoDock parameter file (dpf).

Adjustment of Atomic Solvation Parameters. The atomic charges are related to both the electrostatics and desolvation terms, and the latter term in AutoDock4 was developed along the lines of Wesson et al.[44] and Stouten et al.[45] The atomic solvation parameter and the amount of desolvation are required for evaluating the energetic term of desolvation. The atomic solvation parameter (S_i) in AD4 was determined by a simple linear model:

$$S_i = (\mathrm{ASP}_k + \mathrm{QASP} \times |q_i|), \qquad k = \mathrm{C, A, N, O, S, H}$$

where ASP_k and QASP are the intercept and the regression coefficient, respectively, and q_i is the atomic charge. In this work, we adopt the approach of Bikadi et al.[46] to tune the QASP values for different atomic charge models and retain other calibrated parameters in the original desolvation function of AutoDock4. The new QASP parameters were adopted as 0.006393 and 0.006383 for RP and AP, respectively.

Robust Regression with the FAST-LTS Algorithm. We performed robust regression analyses with the least-trimmed squares (LTS) estimator,[47] which has high-breakdown point, and the influence of the outliers can be mitigated. The computational cost of LTS regression for systems in this study (data set size <200; the number of variables <6) is a few minutes using a single core Xeon X5690 core.

Instead of minimizing the sum square of all residuals of a data set with size n, as in OLS regression, the LTS regression minimizes the sum of squared residuals over a subset of h samples:

$$\sum_{i=1}^{h} (r^2)_{i:n}$$

In calculating the LTS estimator, first, all of the squared residuals r_i's are sorted, and the h smallest squared residuals are selected to calculate the estimator. The absolute residual $|r_i|$ of a sample point i can be considered as its distance to the constructed hyperplane, i.e., the multivariate linear regression model. The detailed analysis of LTS has been described by Rousseeuw et al.[48,49] The FAST-LTS algorithm[49] implemented as "ltsReg" in the "robustbase" package[50] of R (http://www.r-project.org/) was used in this work. The FAST-LTS algorithm starts with randomly selecting p samples, where p is the number of variables in the regression model. Then, a hyperplane (dimension $p - 1$) through these p samples is constructed. The residuals of all n samples are evaluated with respect to the constructed p-subset hyperplane and then sorted. According to the calculated residuals of all the samples, a new subset of h samples with smallest absolute residuals was selected. Subsequently, two C-steps (where C stands for "concentration") are carried out. In a C-step, the ordinary least-squares regression is performed on the h-subset selected in the previous procedure, all the n residuals are evaluated with respect to this regression model and sorted. Only two C-steps are needed because the data size in our system is smaller than 600.[49] This procedure will be repeated 10^8 times, and the 10 models with lowest sum of squares of the h smallest

Table 1. OLS Regression Results of AutoDock4 Scoring Function with Different Charge Models[a]

combinations		coefficients of energetic terms						
ligand	protein	size	W_{desolv}	W_{estat}	$W_{h\text{-}bond}$	W_{tors}	W_{vdW}	RMSE
Gasteiger	Gasteiger	187	0.120	0.142	0.121	0.283	0.172	2.542
AM1-BCC	Amber99SB	187	0.093	0.125	0.077	0.279	0.167	2.523
RESP	Amber99SB	187	0.107	0.132	0.090	0.301	0.167	2.471

[a] All RMSE values are in kcal/mol. *F*-statistics: all p values $<2.2 \times 10^{-16}$.

residuals will be conducted with more C-steps until convergence. The convergence in FAST-LTS algorithm means that the sum of squared residuals over a subset of h samples of $m - 1^{th}$ C-step is the same as the m^{th} C-step. According to the practice of Rousseeuw et al.,[49] m is often below 10. Finally, the model with lowest sum of squares is reported. The entire procedure was repeated twice to confirm that the results are identical.

RESULTS AND DISCUSSION

Ordinary Least-Squares Regression Models with Three Charge Combinations. To understand the inadequacy of using different charge models with the original AutoDock4 scoring function, we first calculated the RMSE between the experimental binding free energy and the binding free energy estimated by the original AD4 scoring function but with the charges calculated with the RESP model. A very large RMSE value, 7.3 kcal/mol, was found, indicating that recalibrating coefficients is indispensable when different charge models are used. If the Gasteiger charge model was used, as in the original AD4 scoring function, a much smaller RMSE of 2.542 kcal/mol was obtained. Because the PDB entries 1sre and 1stp have been identified as outliers in the previous study,[8] the OLS calibration was done with the remaining 187 complexes. However, these two entries were included in the robust regression, where we showed that these two outliers can indeed be identified. With OLS regression, the AD4 scoring functions with RESP and AM1-CC charges for ligands and AMBER Parm99SB charges for proteins yield slightly lower RMSEs, as shown in Table 1.

Progressively Removing the Outliers in OLS Regression Analysis. In the previous section, we performed the OLS regression for the AutoDock4 scoring function to calibrate new coefficients with various charge models. To our knowledge, most, if not all, empirical or semiempirical scoring functions for protein−ligand interactions are constructed by OLS regression. Selection of training data set is always crucial for the OLS regression approach because the influence of outliers is usually very significant. The resistance of OLS to outliers is almost zero, and the fitted model will probably be affected by any arbitrary outlier. In contrast, the robust regression is usually less influenced by outliers.[48]

On the other hand, it may be anticipated that OLS regression could be improved if the samples with large residuals (probable outliers) are removed, which may (wrongly) suggest that OLS with progressive outlier removal can finally generate the same model as the one generated by robust regression. To assess how OLS models evolve by removing the most apparent outliers, we perform an initial OLS regression with the entire data set with N samples. We then remove the sample with largest residual (i.e., the most apparent outlier), and the OLS regression is

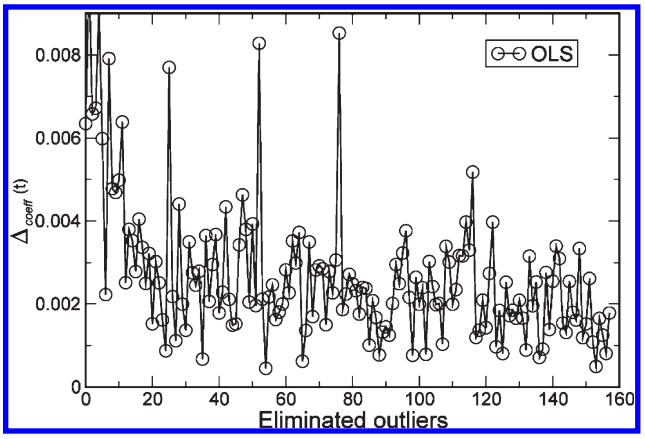


Figure 1. Average of mean coefficient difference $\Delta_{coeff}(t)$ versus number of eliminated outliers. Note that the models are still very unstable even after one-third or more of large residual data points are regarded as outliers and eliminated. The charge combination is RP.

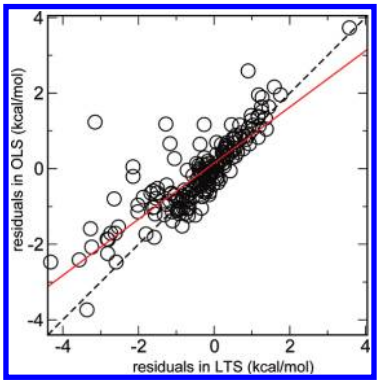


Figure 2. Residual−residual plots between OLS and LTS. The charge combination is RP. The solid red line was obtained by linear fitting between residuals from two regression methods. If no outlier blends with the training set, the residuals of OLS and robust regressions will be similar to the identity line (dashed). The sloped red solid line shows the capacity of resistance to the outliers of different regression models.

performed on the data set with $N - 1$ samples. This so-called evolutionary regression procedure[51] is repeated until the data set size is 30. The coefficients of the N-sample regression model will be compared with the coefficients of the $N - 1$ sample regression model to assess the stability of the models, and the average of mean coefficient difference, $\Delta_{coeff}(t)$ is calculated as follows:

$$\Delta_{coeff}(t) = \sqrt{\frac{1}{5}\sum_{i=1}^{5}(W_{i,t} - W_{i,t-1})^2}$$

$W_{i,t}$ represents the coefficient of the i^{th} energetic term in the t^{th} regression after t samples with the largest residuals are removed. From the curve of $\Delta_{coeff}(t)$ shown in Figure 1, it can clearly be seen that progressively removing the outliers does not lead to stable models with the OLS regression analysis. Note that with LTS robust regression we simply obtain a straight line $\Delta_{coeff}(t) = 0$ for the number of eliminated outliers less than $N - h$ (i.e., 0−97).

Difference Between OLS and LTS Regression Analysis. To illustrate the difference between OLS and LTS regression analysis,

the residual−residual plot (RR plot)[52,53] was made, as shown in Figure 2, which is the scatter plot of the residuals from two regression analyses. The RR plot can be used to characterize the disparity of residuals defined by different methods. Figure 2 indicates that there are indeed significant disparities between the residuals defined by the OLS and LTS methods. It can also be observed that the LTS residuals of most data points with strong disparity are larger. As a result, the outliers possess larger residuals from robust regression but smaller residuals from OLS, which make the solid line tilted. This phenomenon also reflects the capacity of resistance to the outliers of different regression models.

Distribution of Residuals of Three Charge Models. Although robust regression will fit to the majority of data and the models constructed by robust regression are insensitive to outliers, the outliers in the data set still contribute to the RMSE of a model. To reiterate, if there are only a few outliers in a data set, the models (i.e., their coefficients) constructed by robust regression will not be affected, but these outliers will still deteriorate the statistical performance of the models. To fairly assess the statistical performance of a model, it is still important to identify the outliers of a data set. Figure 3 gives the distributions of residuals for robust regression models with three charge combinations. It is interesting to note that the residuals of the model with the AP charge combination give the most symmetric distribution and therefore most "Gaussian-like." The distributions of residuals of the models with the GG and RP charge combinations are rather skewed. It can also be observed that the distribution of residuals of the GG model has a long tail on the left-hand side of the distribution.

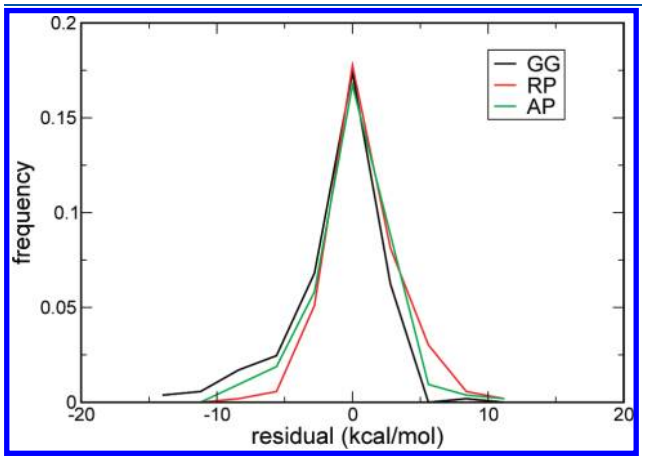


Figure 3. Histograms of residuals based on the model constructed by robust regression analyses. The black, red, and green lines represent residuals distributions of GG, RP, and AP models, respectively.

Identification of Common Outliers to Three Charge Models. A natural strategy to determine the data set for model construction is to remove the common outliers to three charge models. To determine the common outliers, the residuals obtained from the LTS regression were first sorted as shown in Figure 4. To facilitate easy recognition of outliers, a red line that fit the residuals between top 25% and 75% was drawn. A data point that is too far away (larger than the criterion shown below) from the red line in Figure 4 is considered as an outlier. The criterion for the outlier detection here is defined as the absolute value of the y intercept of the red line. It is seen that the GG model possess the largest number of outliers, compared to the AP and RP models. We removed the union of identified outliers from three charge model combinations and finally obtained a common data set of 147 complexes. In the following sections, these 147 complexes will be designated as the "clean" set. The outliers of three robust models are listed in Tables S1, S2, and S3 of Supporting Information.

To obtain the final regression models with three charge combinations, another OLS regression on the clean set was performed, following the previous robust regression procedure.[48] In the following sections, the models constructed by first outlier detection with LTS regression analysis and then OLS regression are called "robust models." Table 2 gives the robust models for the three charge combinations and their RMSE's on the training set (i.e., the clean set). It can be seen that the RAP model (robust model with the charge combination AP) has the lowest RMSE value of 1.637 kcal/mol. It should be noted that the performance of the RGG or RRP model is almost as good as that of the RAP model, which may be due to the fact that the bad data points has

Table 2. Coefficients of the Robust AutoDock4 Scoring Functions[a]

	coefficients of different energetic terms					
models	W_{desolv}	W_{estat}	$W_{h\text{-}bond}$	W_{tors}	W_{vdW}	RMSE
AutoDock4^{RGG}	0.0996	0.0241	0.1806	0.3594	0.1734	1.664
AutoDock4^{RAP}	0.0993	0.0491	0.1565	0.3422	0.1736	1.637
AutoDock4^{RRP}	0.0954	0.0661	0.1521	0.3618	0.1698	1.641

[a] Size of the clean set is 147. All RMSE values are in kcal/mol. F-statistics: all p values $<2.2 \times 10^{-16}$.

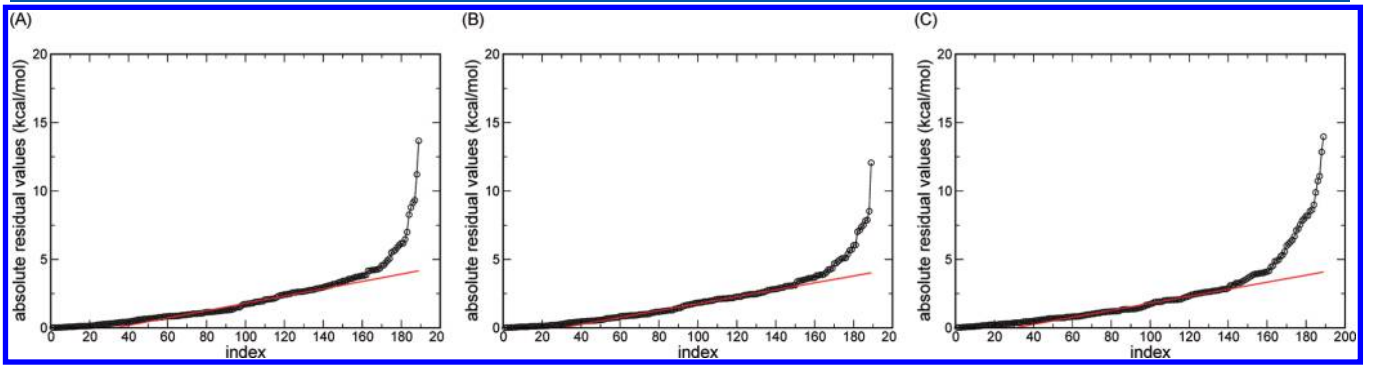


Figure 4. Sorted absolute residuals based on robust regression analysis for the (A) RP and (B) AP and (C) GG charge combinations. The index represents the rank of residuals. Red lines are fitted to the residuals between top 25% and 75%.

Table 3. Cross-Validation of Three Robust Regression Models in This Work[a]

combination	LOO-CV		MCCV	
	S_{PRESSS}	q^2	S_{PRESS}	q^2
AutoDock4^{RGG}	1.732	0.675	1.782	0.657
AutoDock4^{RAP}	1.707	0.684	1.749	0.670
AutoDock4^{RRP}	1.711	0.683	1.755	0.668

[a] All RMSE values and S_{PRESS} are in kcal/mol.

been removed. More detailed comparisons of different charge combinations are discussed in the last subsection of this section.

We further performed two types of cross-validation, the leave-one-out cross-validation (LOO-CV) and leave-group-out cross validation (LGO-CV), shown in Table 3. LOO-CV is a popular approach, but it has recently been discussed for its possible fallacies.[54] Shao demonstrated LOO-CV is asymptotically inconsistent,[55] and Golbraikh showed the high value q^2 of LOO-CV is the necessary but not the sufficient condition for a good QSAR model with high predictive power.[56] Therefore, we also performed the LGO-CV, which is also known as the Monte Carlo cross-validation (MCCV). LGO-CV is conducted by randomly sampling a test set from a group of data points with as many iterations as possible. Based on the suggestions of Konovalov et al.,[23] we divided the clean set into one-half for training and one-half for testing. With 1000 iterations, the average values of S_{PRESS} and q^2 were summarized in Table 3. The S_{PRESS} and q^2 are given by following equations:

$$S_{PRESS} = \sqrt{\frac{\sum_i (E_{i,\mathrm{pred}} - E_{i,\mathrm{exp}})^2}{(N-k)}};$$

$$q^2 = 1 - \frac{\sum_i (E_{i,\mathrm{pred}} - E_{i,\mathrm{exp}})^2}{\sum_i (E_{i,\mathrm{exp}} - E_{i,\mathrm{mean}})^2}$$

E_{pred} and E_{exp} are the predicted and experimental binding free energies, respectively. E_{mean} is the mean value of experimental binding free energy of all observed cases. N is the size of the training set. The degree of freedom, k, is 5 for all the regression models in this study. The results shown in Table 3 indicate that all assessments of cross-validation are comparable for the performance on the training set. It was shown that the RAP and RRP models gave slightly smaller prediction errors and higher correlations, compared to the RGG model, but the differences in the numerical values of this statistical assessment may not be significant.

Assessment with External Complexes. To assess whether the performances of our new robust models are sensitive to the data set, a benchmark on an external data set of protein−ligand complexes from PDBbind[57,58] was conducted. PDBbind is currently the largest public database that contains the structural information and binding affinities of receptors and ligands. We started with the data set with 1741 protein−ligand complexes from the 2009 version of PDBbind, which is the so-called "refined set." In the PDBbind refined set, ligands with added hydrogens and Gasteiger charges have been prepared, and receptors structural files are arranged as biological assemblies. We first filtered out 211 complexes whose net charges of ligands are not consistent with the calculation with the "estimateFormalCharge" function of Chimera. In addition to these ligands with problematic net charges and protonated states, some complexes have problems in atomic charges or energy calculations. For example, MOPAC or GAUSSIAN could not be used to calculate too large molecules within an acceptable time, autogrid4 cannot generate maps for the molecule with more than 32 769 atoms, and autodock4 cannot calculate the energies of a ligand with more than 32 torsions. These complexes were further removed. Finally, 1427 complexes from the PDBbind refined set was used as the test set in this study.

Table 4. Performance of the Robust AutoDock4 Scoring Functions and Two Other Recent Scoring Functions Tested with the PDBbind Data Sets[a]

scoring function	N_{train}	N_{test}	R_P	R_S	SD	ME
AutoDock4^{RGG}	147	1427	0.604	0.615	1.61	1.26
AutoDock4^{RAP}	147	1427	0.606	0.617	1.60	1.25
AutoDock4^{RRP}	147	1427	0.595	0.610	1.62	1.26
original AutoDock4^{GG}	187	1427	0.562	0.594	1.66	1.31
sfc_290m	290	919	0.492	0.555		
sfc_229m	229	919	0.501	0.558		
sfc_frag	130	919	0.525	0.576		
PDSE-SVM	278	977	0.517	0.535	1.84	1.42

[a] SD and ME are presented in the pK_d unit. The binding free energy in kcal/mol at 298 K was converted to the pK_d unit by dividing with the factor of −1.36. F-statistics: all p values $<2.2 \times 10^{-16}$.

Table 4 shows the statistical performances of the three robust AutoDock4 scoring functions and two other recent protein−ligand scoring models, SFCscore[4] and PDSE-SVM,[5] on PDB bind data sets. The three robust AutoDock4 scoring functions have significantly higher correlations [Pearson's correlation coefficient (R_P) and Spearman's correlation coefficient (R_S)] as well as smaller standard deviations (SD) and mean errors (ME). For the comparison with SFCscore, we only showed the results of their models that were constructed by multivariate linear regression. Because the test set we used (PDBbind v2009) and the test set of PDSE-SVM (PDBbind v2005) have an overlap of only 634 complexes, we also made an assessment by using the refined set of PDBbind version 2005. Our robust AutoDock4 scoring functions gave comparable results (R_P = 0.540−0.578, R_S = 0.553−0.601) to the performance of PDSE-SVM. Our assessment indicated that AutoDock4^{RGG} gave slightly better statistics than AutoDock4^{RRP}. However, the small difference in the numerical values of statistics may not be significant.

Assessment of Binding Pose Prediction with External Decoys. The performance of binding pose prediction of the three robust AutoDock4 scoring functions and the original AutoDock4 GG model was assessed by the decoy sets of 100 protein−ligand complexes from Wang et al.[25] In this test, 100 ligand conformations near the binding site in each complex were generated by using AutoDock3, and the native ligand conformation of each complex was also included. All structural information with hydrogens added for ligands and receptors is available.[59] After the atomic charges were calculated, we performed local minimizations to optimize too close contacts of original structures (only for native ligand conformations) in the same procedure as the preparation of the LPDB training data set. The mean values of RMSD's between the original and minimized conformation are 0.57, 0.66, and 0.74 Å for RGG, RAP, and RRP, respectively.

Table 5. Success Rates of Binding Site Prediction by Different Scoring Functions[a]

	success rate (%) for different RMSD criteria				
scoring function	≤1 Å	≤1.5 Å	≤2 Å	≤2.5 Å	≤3 Å
DrugScore^{CSD}	83	85	87		
AutoDock4^{RAP}	83	85	87	87	87
AutoDock4^{RGG}	80	82	86	86	86
AutoDock4^{RRP}	79	81	84	85	85
original AutoDock4^{GG}	74	76	79	79	79
Cerius2/PLP	63	69	76	79	80
SYBYL/F-Score	56	66	74	77	77
Cerius2/LigScore	64	68	74	75	76
DrugScore	63	68	72	74	74
Cerius2/LUDI	43	55	67	67	67
X-Score	37	54	66	72	74
AutoDock3	34	52	62	68	72
Cerius2/PMF	40	46	52	54	57
SYBYL/G-Score	24	32	42	49	56
SYBYL/ChemScore	12	26	35	37	40
SYBYL/D-Score	8	16	26	30	41

[a] Except for the results of the AutoDock4 scoring functions, the results of DrugScore^{CSD} and other scoring functions were taken from Velec et al.[26] and Wang et al.,[25] respectively. Scoring functions are sorted by the number of cases under 2 Å.

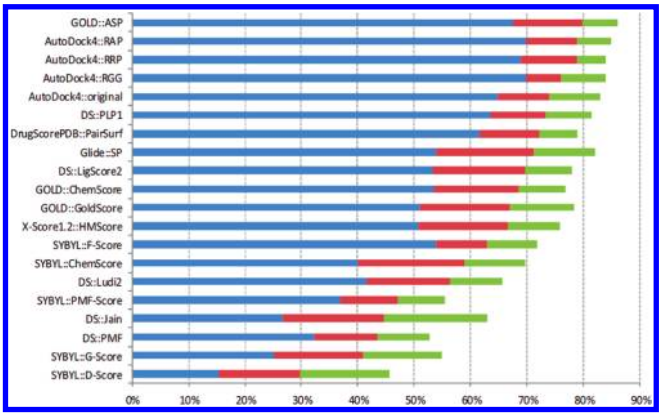


Figure 5. Comparison of the success rates of AutoDock4 scoring functions and 16 scoring functions provided by Cheng et al.[28] The cutoffs are RMSD < 1.0, < 2.0, and < 3.0 Å (blue, red, and green bars), respectively. The native binding poses of ligands were included in the decoy sets. Scoring functions are sorted by the number of cases under 2 Å.

Ideally, if a scoring function could recognize near-native structures among a set of decoys, at least some of the near native conformations (i.e., with small RMSD's with respect to the native ligand conformation) should have the best scores or the lowest predicted binding free energies. Thus, each conformation with lowest predicted energy and corresponding RMSD with respect to the native conformation was recorded. The success rate can be defined according to different criteria, as shown in Table 5, which gives success rates of AutoDock4 scoring functions and other scoring functions. We found that AutoDock4 made a remarkable improvement compared to AutoDock3 on the same decoy set.

Table 6. Success Rates of Binding Pose Prediction of Various Scoring Functions on Three Classes of Complexes[a]

	success rate (%, RMSD ≤ 2 Å)			
	overall	hydrophilic	mixed	hydrophobic
scoring function	(100)	(44)	(32)	(24)
AutoDock4^{RAP}	87	89	91	79
AutoDock4^{RGG}	86	86	91	79
AutoDock4^{RRP}	84	84	91	75
original AutoDock4^{GG}	79	77	81	79
Cerius2/PLP	76	77	78	71
SYBYL/F-Score	74	75	75	71
Cerius2/LigScore	74	77	75	67
DrugScore^{PDB}	72	73	81	58
Cerius2/LUDI	67	75	66	54
X-Score	66	82	59	46
AutoDock3	62	73	53	54
Cerius2/PMF	52	68	44	33
SYBYL/G-Score	42	55	34	29
SYBYL/ChemScore	35	32	34	42
SYBYL/D-Score	26	23	28	29

[a] Scoring function success rates were adopted from Wang et al.[25] except for AutoDock4 scoring functions. Scoring functions are sorted according to the overall success rates.

The RAP model can even achieve the same success rates of DrugScore^{CSD}. The RAP and RRP models are not identical, although AM1-BCC is a semiempirical quantum mechanical model that aimed to reproduce the RESP results as much as possible. On the other hand, we should stress that the performance also strongly depends on the test set. In 2009, Cheng et al.[28] published a comparative assessment of scoring functions on a new decoy set, which consists of 195 complexes with reassessed quality of structures, binding data, and components of protein complexes. The ligand conformations were obtained from four docking software packages, with the aim to reduce the bias in binding pose selection. Figure 5 gives the comparison of the success rates of AutoDock4 scoring functions and 16 scoring functions. The robust AutoDock4 functions achieve excellent success rates compared to most of other scoring functions, as shown in Figure 5. To assess whether our robust functions exhibit strong class dependence, we delineate the success rate results into three classes (hydrophilic, hydrophobic, and mixed). Table 6 summarizes the success rates in these three classes of 100 complexes.[25] The robust AutoDock4 functions achieve outstanding success rates in all three classes. It is noted that there is no difference in the accuracy of binding pose prediction between using the RAP charge model and the original AutoDock4 scoring function for the hydrophobic class of complexes, but there is a significant difference for the hydrophilic class. To further investigate the potential reasons for such observed differences, we inspected the cases that were predicted differently by these two scoring functions, and found that the four D-xylose isomerase complexes (8xia, 4xia, 2xia, and 2xis) in hydrophilic class could be successfully predicted by RAP but not by the original AutoDock4 scoring function. These cases have some common features: two metals (magnesium or manganese) and a D-xylose in the active site. The difference between the estimated

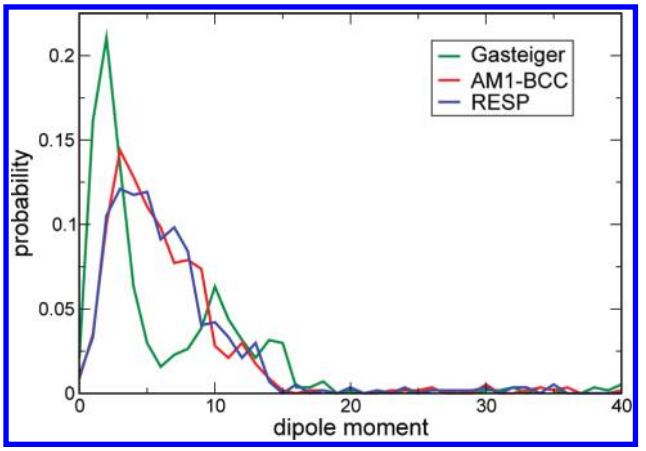


Figure 6. Distributions of dipole moments (in D) of 569 neutral ligands. The values of dipole moment were calculated by three charge models.

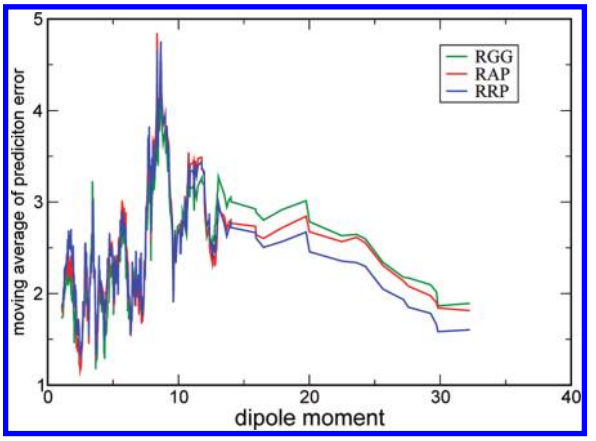


Figure 7. Moving average of prediction errors (in kcal/mol) versus dipole moments (in D). The prediction error was the deviation between estimated and experimental binding free energy. The values of dipole moment were sorted according to the dipoles calculated with the RESP charge model.

energies of the native poses of D-xylose from RAP and the original AutoDock4 scoring function is mainly due to the electrostatic energetic term (~0.7 kcal/mol) between an oxygen atom of ligand and a metal on the protein site. The different charges of the oxygen atom result in different electrostatic interaction.

Performance of Three Models for Large Dipole Moment Cases. So far, our assessments indicate that the robust model with the GG charge combination (Gasteiger models for both ligand and protein) can achieve similar statistical performances of the robust models with the other two charge combinations, in which quantum chemical calculations need to be carried out. One possible explanation for such close performances could be attributed to the heterogeneity of the data set. Regression analysis, when properly performed, provides the suitable (and subtle) balances among different energetic terms, and the shortcoming or the inaccuracy of some energetic terms can be mitigated by reducing their weighting coefficients. It is therefore worthwhile to assess the three robust models with different charge combinations in the subset of the test set, where the differences of the charge models could be most pronounced. Because different charge models (Gasteiger, RESP, AM1-BCC) mainly affect the distributions of the partial charges of the molecules, not the total charge of the molecules, we should be able to see the differences of the charge models on the subset in which ligands are neutral. On such subset of complexes with neutral ligands, it is especially of interest to see the dependence of statistical performance on the dipole moments of ligands, because the dipole moments give the largest contribution to the electrostatic energies for the neutral ligands.

Table 7. RMS Prediction Errors of AutoDock4 Scoring Functions for Neutral Ligands in the Testing Set

scoring function	569 cases (kcal/mol)	43 cases[a] (kcal/mol)
AutoDock4^{RGG}	3.004	3.022
AutoDock4^{RAP}	3.087	2.755
AutoDock4^{RRP}	3.088	2.702
original AutoDock4^{GG}	3.253	2.938

[a] These 43 cases are large dipole moment ligands. (>12.5 D).

For the 569 neutral ligands of the PDBbind test set, the dipole moment distributions according to three kinds of charge models are shown in Figure 6. One can easily recognize the differences in the distributions of the dipole moments calculated with three different charge models. We further sorted the prediction errors according to the dipole moments and took moving average to smooth out the large fluctuation for clearer visualization of their tendencies, shown in Figure 7. We can see pronounced differences of prediction errors for the cases with large dipole moments (larger than 12.5 D, from Figure 6). In Table 7, according to the RMS prediction errors it was clearly indicated that RRP has the best statistical performance for the subset of complexes with neutral ligands having large dipole moments. We also analyzed the dipole moment distributions and prediction errors in our training set, whose size may in turn be too small to provide significant statistical differences. The results are given in Supporting Information (Figures S1 and S2 and Table S4).

CONCLUSION

We have constructed three robust protein−ligand free energy models for three popular charge combinations. The combination of AM1-BCC or RESP charges for ligands and Amber99SB charges for proteins performs statistically better than the combination of Gasteiger charges for ligands and proteins does. Our results also indicate that the use of more advanced charge models may lead to more accurate estimates of protein−ligand binding free energy, especially for the protein−ligand complexes with large dipole moment neutral ligands.

Nevertheless, construction of free energy models (or scoring functions) for protein−ligand interactions based on regression analysis remains a challenging task. There are many uncertainties in the experimental information and in the preparation of protein and ligand files, e.g., determination of protonation states, number of rotatable bonds, etc. In this work, the flexibility of the protein−ligand complex was not yet explicitly taken into account for constructing the scoring functions. The contribution of stable water molecules in the protein binding pocket was also not yet included. The metals were only treated in the classical manner, and the consideration of the entropy contribution is certainly incomplete. Yet, with robust regression, we were able to show the five energy terms in the AutoDock4 scoring functions can capture the essential picture of the protein−ligand interactions. It should be stressed that the same scoring function was applied

in AutoDock4 to both binding pose prediction and binding free energy evaluation, and its molecular mechanics-based semiempirical nature allows sensitive recognition of protein conformational changes. This is not the case for many protein−ligand scoring functions that adopt relatively coarse-grained potential or crude distance criteria, e.g. XSCORE, ChemScore, PLP, etc.

Our analyses for the performances of different scoring functions on the subset of neutral ligand may indicate that the accuracy of such regression models may be improved still further when the protein−ligand complexes are suitably classified. However, this also implies that a larger training set is needed if multivariate regression is to be applied for different protein−ligand interaction classes. Construction of larger databases with structural and binding affinity information of protein−ligand complexes, similar to the endeavor of PDBbind, is indispensable for establishing such two-staged (first class identification, model selection, and then free energy prediction) free energy models.

■ ASSOCIATED CONTENT

Supporting Information. Outliers detected by the robust regression analysis in three charge model combinations were listed. The plots of dipole moment distributions and prediction errors in the training set were provided. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: jlin@ntu.edu.tw. Telephone: +8862-27823212 ext 886.

■ ACKNOWLEDGMENT

J.-H.L. was funded by National Science Council of Taiwan grant no. NSC 98-2627-B-001-002, 98-2323-B-002-001, 98-2323-B-077-001, 97-2323-B-002-015, and 97-2923-M-001-001-MY3. Supported from Research Center for Applied Sciences, Academia Sinica was also greatly acknowledged. The authors gratefully acknowledge Dr. Robert O. Jones in Forschungszentrum Jülich for his comments on the manuscript.

■ REFERENCES

(1) Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(2) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **1997**, *72*, 1047–1069.

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

(4) Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFCscore: Scoring functions for affinity prediction of protein-ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 395–419.

(5) Das, S.; Krein, M. P.; Breneman, C. M. Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model.* **2010**, *50*, 298–308.

(6) Kramer, C.; Gedeck, P. Global Free Energy Scoring Functions Based on Distance-Dependent Atom-Type Pair Descriptors. *J. Chem. Inf. Model.* **2011**, *51*, 707–720.

(7) Hansch, C.; Maloney, P. P.; Fujita, T. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178–180.

(8) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(9) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–1152.

(10) Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(11) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

(12) Lin, J. H. Accommodating protein flexibility for structure-based drug design. *Curr. Top. Med. Chem.* **2011**, *11*, 171–178.

(13) Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632–5633.

(14) Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **2003**, *68*, 47–62.

(15) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(16) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(17) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.

(18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(19) Dreizler, H. a. D. G. Rotation spectrum, r_o structure, and dipole moment of dimethylsulfoxide. *Z. Naturforsch.* **1964**, *19a*, 512–514.

(20) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26*, 915–931.

(21) Cho, A. E.; Rinaldo, D. Extension of QM/MM docking and its applications to metalloproteins. *J. Comput. Chem.* **2009**, *30*, 2609–2616.

(22) Tsai, K. C.; Wang, S. H.; Hsiao, N. W.; Li, M.; Wang, B. The effect of different electrostatic potentials on docking accuracy: A case study using DOCK5.4. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3509–3512.

(23) Konovalov, D. A.; Llewellyn, L. E.; Heyden, Y. V.; Coomans, D. Robust cross-validation of linear regression qsar models. *J. Chem. Inf. Model.* **2008**, *48*, 2081–2094.

(24) Wang, J. C.; Lin, J. H. Robust regression analysis of protein-ligand binding free energy models: toward the identification of druggable genomes. *Int. J. Syst. Syn. Biol.* **2010**, *1*, 339–354.

(25) Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.

(26) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296–6303.

(27) Xie, Z. R.; Hwang, M. J. An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinf.* **2010**, *11*, 298.

(28) Cheng, T. J.; Li, X.; Li, Y.; Liu, Z. H.; Wang, R. X. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

(29) Roche, O.; Kiyama, R.; Brooks, C. L. Ligand-Protein DataBase: Linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592–3598.

(30) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

(31) Mehler, E. L.; Solmajer, T. Electrostatic effects in proteins-comparison of dielectric and charge models. *Protein Eng.* **1991**, *4*, 903–910.

(32) Bohm, H. J. 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.

(33) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the resp model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(34) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.

(36) Hehre, W. J.; Ditchfie., R; Pople, J. A. Self-consistent molecular-orbital Methods. XII. Further extensions of gaussian-type basis sets for use in molecular-orbital studies of organic-molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(37) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(38) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(39) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(40) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. Blue Obelisk-Interoperability in chemical informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.

(41) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(42) Autenrieth, F.; Tajkhorshid, E.; Baudry, J.; Luthey-Schulten, Z. Classical force field parameters for the heme prosthetic group of cytochrome c. *J. Comput. Chem.* **2004**, *25*, 1613–1622.

(43) Oda, A.; Yamaotsu, N.; Hirono, S. New AMBER force field parameters of heme iron for cytochrome P450s determined by quantum chemical calculations of simplified models. *J. Comput. Chem.* **2005**, *26*, 818–826.

(44) Wesson, L.; Eisenberg, D. Atomic Solvation Parameters Applied to Molecular-Dynamics of Proteins in Solution. *Protein Sci.* **1992**, *1*, 227–235.

(45) Stouten, P. F. W.; Frommel, C.; Nakamura, H.; Sander, C. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. *Mol. Simul.* **1993**, *10*, 97–120.

(46) Bikadi, Z.; Hazai, E. Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *J. Cheminf.* **2009**, *1*, 15.

(47) Rousseeuw, P. J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880.

(48) Rousseeuw, P. J.; Leroy, A. M. In *Robust Regression and Outlier Detection*; Barnett, V. et al. , Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 1987; pp 9−17, 112−142.

(49) Rousseeuw, P. J.; Van Driessen, K. Computing LTS regression for large data sets. *Data Min. Knowl. Dis.* **2006**, *12*, 29–45.

(50) Rousseeuw, P.; Croux, C.; Todorov, V.; Ruckstuhl, A.; Salibian-Barrera, M.; Verbeke, T.; Maechler, M. *robustbase: Basic Robust Statistics*, R package version 0.7-6; Swiss Federal Institute of Technology Zurich: Zurich, Switzerland; http://CRAN.R-project.org/package=robustbase, (accessed August 11, 2011).

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

(52) Anderson, R. In *Modern Methods for Robust Regression*; Liao, T. F., Ed.; SAGE: Thousand Oaks, CA, 2008; Chapter 4, pp 67−68.

(53) Tukey, J. W. Graphical displays for alternative regression fits. In *Robust Statistics and Diagnostics*, Part 2; Stahel, W., Weisberg, S., Eds.; Springer-Verlag: New York, 1991; pp 309.

(54) Hawkins, D. M.; Kraker, J. Deterministic fallacies and model validation. *J. Chemom.* **2010**, *24*, 188–193.

(55) Shao, J. Linear-Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.

(56) Golbraikh, A.; Tropsha, A. Beware of q(2)!. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(57) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(58) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(59) Wang, R.; Fang, X. *X-SCORE*; Department of Internal Medicine, University of Michigan Medical School: Ann Arbor, MI; http://sw16.im.med.umich.edu/software/xtool/, (accessed April 28, 2011).