JCTC Journal of Chemical Theory and Computation

# Mixed Quantum Mechanics/Molecular Mechanics Scoring Function To Predict Protein−Ligand Binding Affinity

Seth A. Hayik,[†,‡] Roland Dunbrack, Jr.,[†] and Kenneth M. Merz, Jr.*[,‡]

*Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111, and Department of Chemistry, Quantum Theory Project, University of Florida, P.O. Box 118435, Gainesville, Florida 32611*

Received June 8, 2010

**Abstract:** Computational methods for predicting protein−ligand binding free energy continue to be popular as a potential cost-cutting method in the drug discovery process. However, accurate predictions are often difficult to make as estimates must be made for certain electronic and entropic terms in conventional force field based scoring functions. Mixed quantum mechanics/molecular mechanics (QM/MM) methods allow electronic effects for a small region of the protein to be calculated, treating the remaining atoms as a fixed charge background for the active site. Such a semiempirical QM/MM scoring function has been implemented in AMBER using the DivCon program and tested on a set of 23 metalloprotein−ligand complexes, where QM/MM methods provide a particular advantage in the modeling of the metal ion. The binding affinity of this set of proteins can be calculated with an $R^2$ of 0.64 and a standard deviation of 1.88 kcal/mol without fitting and an $R^2$ of 0.71 and a standard deviation of 1.69 kcal/mol with fitted weighting of the individual scoring terms. In this study we explore the use of various methods to calculate terms in the binding free energy equation, including entropy estimates and minimization standards. From these studies we found that using the rotational bond estimate of ligand entropy results in a reasonable $R^2$ of 0.63 without fitting. We also found that using the ESCF energy of the proteins without minimization resulted in an $R^2$ of 0.57, when using the rotatable bond entropy estimate.

## Introduction

It is important to accurately model ligand interactions so that computational screening can be effectively applied to lower the cost and time involved in drug discovery.[1,2] Protein−ligand interactions involve a complicated mixture of electrostatic, dispersion, and other interactions and therefore represent a challenge to model and predict accurately. Methods for predicting ligand affinity vary in the scoring function and structure prediction methods used, and each has different advantages and disadvantages.[3–5] Often, these schemes sacrifice accuracy to gain speed or improve accuracy at greater computational expense but then are too complicated and slow to be used for large-scale applications.[6,7] For these reasons research in this area continues to thrive, searching for a model that achieves a good compromise between speed and accuracy to predict binding affinities and score binding poses.

Empirical scoring functions are composed of several terms trained on experimental binding data to generate general parameters.[8–10] These functions may perform well and are quick, but have limitations due to the way these scoring functions are derived and the training sets used to construct them. The accuracy of empirical and knowledge-based scoring functions is dependent on the size and variety of the training set used. These functions may fail if the systems being examined are too different from any ligands in the training set.[11] This leaves the scoring function without knowledge of the systems being examined, and therefore, it essentially must draw conclusions from estimates of other ligands.

* Corresponding author phone: (352) 392-6973; fax: (352) 392-8722; e-mail: merz@qtp.ufl.edu.

[†] Fox Chase Cancer Center.

[‡] University of Florida.

10.1021/ct100315g 
Published on Web 09/07/2010

Another class of scoring functions is constructed using a potential based on classical molecular mechanics force fields.[12–16] These methods either exclude electronic effects or account for them with an empirical parameter that may not always apply. These methods also have problems accurately predicting properties for nonstandard residues and metalloenzymes, which are difficult to describe with classical methods,[17] although some successful examples have been reported.[18,19]

Quantum mechanics (QM) methods have begun to demonstrate their usefulness as scoring functions for calculating ligand binding free energies as computer power has increased. Until recently, QM methods were primarily used only for smaller systems because the cost associated with these methods was untenable for large-scale screening. Efforts have been made to decrease the time required for QM calculations of proteins using various methods, making them more viable in scoring functions.[20–24] We recently developed QMScore,[25–27] a scoring function using a full QM potential that uses a linear scaling semiempirical method within the program DivCon[22,23] to calculate properties of the protein, ligand, and complex and to combine them into a scoring function to give a binding free energy.

QM methods have also been incorporated into molecular mechanics programs[28] and have seen a resurgence recently as computing power has increased to create mixed quantum mechanics/molecular mechanics (QM/MM) methods.[29] These methods have proven quite successful for molecular dynamics simulations and reaction mechanism studies in biological systems.[30–33] QM/MM methods allow a small region of interest to be explored in more detail while treating the surroundings with a faster method to save computational time. Some recent structure-based drug design methods have used quantum mechanics,[34,35] QM/MM,[36–40] or QM/QM[41] methods to calculate protein−ligand binding. These methods take advantage of the mixed method's ability to specify a region of interest, usually the ligand, with an expensive Hamiltonian while treating the remaining system with a cheaper Hamiltonian to get more accuracy for the ligand and its surroundings while not becoming prohibitively expensive.

It has been shown that polarization[42–45] and charge transfer effects[26,27] can play an important role in docking a ligand to a target, effects that classical methods often cannot take into account. In particular, QM methods have shown their potential in predicting binding affinities for metalloenzymes. Difficulties with metalloenzymes arise from the expanded valence and the high charge of the metal atom and the charge transfer associated with ligand binding.[19] Our previously developed method, QMScore,[25–27] allows the entire protein to be calculated using a semiempirical QM potential, with the drawback that it is time-consuming to calculate the entire protein at a QM level. However, the benefit is that this method allows electronic effects such as charge transfer to be captured throughout the protein.

QM/MM methods allow polarization effects near the ligand to be taken into account, as QMScore does, without the need to include the entire protein in the quantum region. While long-range polarization may impact ligand binding slightly, Illingworth et al. found that a majority of the polarization energy is within 5 Å of the ligand and polarization's effect on the charges was fairly short-ranged when examining MM charge polarization.[46] Including the first shell of residues allows this polarization to be fully incorporated into the energy terms of the scoring function. These QM/MM methods also have an advantage over many other methods in that specific parameters are not necessary for the ligand or many common metals that may be present.

In this paper, we develop a scoring function comprising several terms: an interaction energy derived from QM/MM energies calculated with AMBER 10 with DivCon, a solvation term based on the change in surface area upon binding, and entropic terms based on the QM/MM frequencies of the system or the rotatable bonds in the ligand to determine the overall binding free energy. We compare the results of the new scoring function with the full QM calculations of Raha and Merz, focusing on a set of zinc metalloenzymes previously modeled with QMScore in a full QM treatment.[25] Our results show that the QM/MM-based scoring function can be optimized to perform nearly as well as the full QM calculation. In addition, we examine a number of factors of the calculation that can be altered to explore accuracy/speed trade-offs for this QM/MM method, such as the number of minimizations and the entropy calculation method.

## Methods

**QM/MM Implementation.** The linear scaling program DivCon,[22,23] integrated into AMBER 10, was used to model the QM region around the ligand. This combined (active site/ligand) region was included in the semiempirical QM calculation, while atoms outside of this region were treated with the classical AMBER ff99SB[47] force field. This gives the effective Hamiltonian of the system, which is a sum of the MM ($\hat{H}_{MM}$), QM ($\hat{H}_{QM}$), and QM/MM interaction ($\hat{H}_{QM/MM}$) Hamiltonians:

$$\hat{H}_{eff} = \hat{H}_{MM} + \hat{H}_{QM} + \hat{H}_{QM/MM} \quad (1)$$

Splitting the system into regions like this leaves many covalent bonds between the two regions severed, resulting in dangling bonds and charge imbalances in both regions. This is addressed by adding a hydrogen link atom to the QM atom in the broken bond that is formed, similar to the implementation in Dynamo.[48] The link atom is forced to lie along the bond vector so that no extra degrees of freedom are introduced to the system, and it is treated like a regular QM atom throughout the calculation. The forces on the link atom are then distributed to the QM and MM atoms that make up the bonding pair, and any interactions involving the MM atom are treated classically.[49]

The electrostatics of the link atom must be carefully considered to avoid false polarization of the QM boundary atom and to maintain a constant charge for the QM region. This is accomplished by spreading the charge of the QM region removed from $\hat{H}_{MM}$ across all of the MM atoms in the system at the beginning of a run. This initial setup step adds a slight cost to the QM/MM calculation, but is not prohibitively expensive as it is a one-time calculation. To

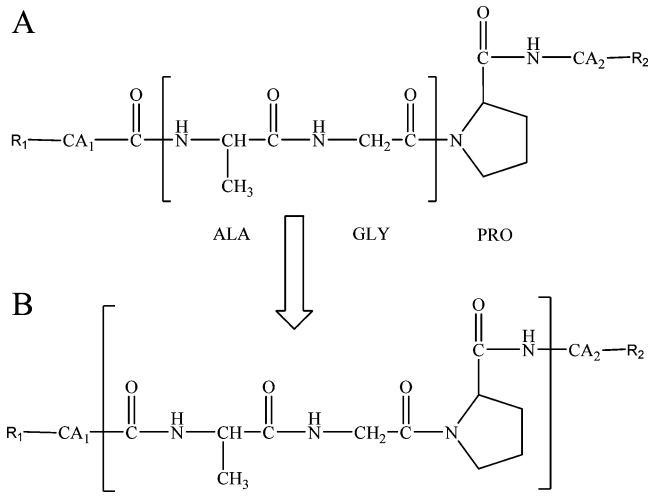


***Figure 1.*** Graphical representation of the QM/MM partitioning scheme. The area in brackets represents the QM region. (A) The rough QM region with a PRO residue at the boundary cuts the two peptide bonds between the C and N. (B) The refined structure extends the QM boundary to include the PRO residue and moves the boundary to cut the C−CA bond to avoid spurious polarization at the boundary.

avoid overpolarization of the QM/MM boundary, the classical atom involved in the link atom bond is ignored by the link atom, and the van der Waals interactions of the bonding pair are treated classically. This results in more stable charge distributions for the atoms in the QM region[49] and eliminates any false, highly repulsive forces between the link atom and the MM bonding atom.

The QM/MM method in AMBER allows for the minimization of a system of interest using the conjugate gradient or steepest descent methods. In a QM/MM calculation the forces of the QM region are calculated by the QM program being used and are then transferred and added to the MM forces on the QM atoms, while the MM forces are calculated by AMBER. This allows a fairly simple minimization to be undertaken for a QM/MM calculation while still removing the need to calculate parameters for an organic molecule or metal ion and taking advantage of the parameters available in the force field for defined atom types.

**Preparation of QM/MM Calculations.** In a QM/MM calculation, the boundary must be carefully chosen so that the approximations and assumptions made in a QM/MM system do not significantly alter the accuracy of the calculation. The QM region should not be too large to realize the cost savings with the MM portion of the energy function. In this study the QM region was selected to include not only the ligand, but also the first shell of the active site to capture electronic changes from binding. Any protein residue with at least one atom within 5 Å of the zinc ion was included in the QM region. The QM region was expanded so that highly polarizable peptide bonds were not cut by the boundary as shown in Figure 1. If a proline was present, the entire residue was included along with its peptide bond. This accommodates the QM cut so that it is not adjacent to the carbonyl group of the peptide bond, providing more distance between the polar region and the QM/MM boundary. If any disulfide bond was included in the QM region, the region was

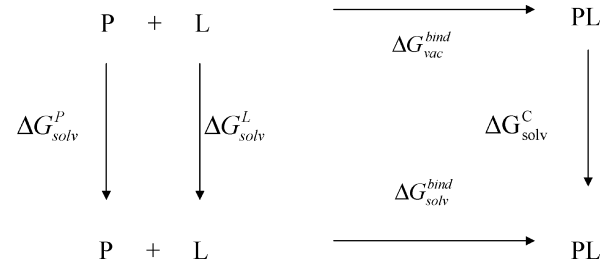


***Figure 2.*** Schematic view of the thermodynamic cycle to calculate binding affinity. The cycle calculates the protein (P), ligand (L), and complex (C) in vacuum and then transfers them to solvent to find the solvation free energy.

expanded to include both partners. Once the QM region was properly defined, the charge of the QM region was calculated on the basis of what residues were present and the charge of the ligand. The active site was visually inspected to confirm the accuracy of the charge.

**Preparation of Proteins.** The structures used in this study were downloaded from the Protein Data Bank (PDB).[50] Protons were added to the heavy atoms using the LEAP module in AMBER 5.0 based on standard geometries and physiological pH, and energy minimization was performed to relax the added protons. Each of these complexes contained a zinc ion in the active site, and a zinc-bound water molecule was added to the uncomplexed proteins to fill the exposed valence. These structures were used as the starting point in this study.

From this point the structures were split into the QM and MM regions for the QM/MM calculations as described above and minimized for 500 cycles using steepest descent and a maximum of another 1500 cycles using the conjugate gradient method in version 10 of AMBER.[51] The long-range cutoff for both the QM and MM regions was set to 100 Å to include long-range effects in the binding affinity calculations, unless otherwise stated. After the initial structures were minimized, the solvation free energies of the protein, ligand, and complex were calculated in DivCon. In addition to the solvation free energy, the vibrational frequencies of the ligand and protein alone and in complex were calculated using DivCon, along with a count of the rotatable bonds on the ligand, to provide an estimate of the vibrational entropy change associated with binding. The convergence criterion for the vibrational entropy step was increased to $1 \times 10^{-6}$ for the changes in the energy and density matrix, and the long-range cutoff was again set to 100 Å.

**Binding Affinity Calculation.** The binding affinity of a protein−ligand complex can be determined using the thermodynamic cycle shown in Figure 2. The quantity of interest is $\Delta G^{\mathrm{sol}}_{\mathrm{bind}}$, the free energy change of binding in solvent, which can be calculated via the alternative path of first desolvating the protein and the ligand, binding of the protein and ligand into a complex, and then solvating the complex. If the free energy change on solvation for each entity $X$ is expressed as $\Delta G^{X}_{\mathrm{solv}}$, then the desolvation contributes $-\Delta G^{\mathrm{protein}}_{\mathrm{solv}}$ and $-\Delta G^{\mathrm{ligand}}_{\mathrm{solv}}$ to $\Delta G^{\mathrm{sol}}_{\mathrm{bind}}$, while the solvation of the complex contributes $+\Delta G^{\mathrm{complex}}_{\mathrm{solv}}$. The free energy of binding in solution can therefore be broken down into the gas-phase binding free energy and the solvation free energy:

$$\Delta G_{\text{bind}}^{\text{sol}} = \Delta G_{\text{bind}}^{\text{gas}} + \Delta\Delta G_{\text{solv}} \quad (2)$$

where

$$\Delta\Delta G_{\text{solv}} = \Delta G_{\text{solv}}^{\text{complex}} - (\Delta G_{\text{solv}}^{\text{protein}} + \Delta G_{\text{solv}}^{\text{ligand}}) \quad (3)$$

The gas-phase free energy of binding contains an enthalpic term from the electrostatic and nonpolar interaction energies and an entropic term from the degrees of freedom of the individual systems:

$$\Delta G_{\text{bind}}^{\text{gas}} = \Delta H_{\text{bind}}^{\text{gas}} - T\Delta S_{\text{bind}}^{\text{gas}} \quad (4)$$

$$\Delta S_{\text{bind}}^{\text{gas}} = \Delta S_{\text{SASA}} + \Delta S_{\text{vib}} \quad (5)$$

where $\Delta S_{\text{SASA}}$ is the entropy change from the solvent-accessible surface area (SASA) and is discussed in further detail below.

The enthalpy term is calculated from the SCF energy in DivCon at the AM1[52] semiempirical level, which was found to work well previously:[41]

$$\Delta H_{\text{bind}}^{\text{gas}} = \Delta H_{\text{f}}^{\text{complex}} - \Delta H_{\text{f}}^{\text{protein}} - \Delta H_{\text{f}}^{\text{ligand}} \quad (6)$$

where

$$\Delta H_{\text{f}} = E_{\text{elec}} + E_{\text{core-core}} + \sum_{\text{atoms}} \Delta H_{\text{f}} \quad (7)$$

$E_{\text{elec}}$ is the electronic energy, $E_{\text{core-core}}$ is the core−core repulsion, and the final term is the sum of the heat of formation for all the atoms. These terms are all calculated within DivCon.

Another important consideration for protein−ligand binding is the change in entropy of the ligand upon binding in the gas phase, $\Delta S_{\text{bind}}^{\text{gas}}$, calculated here as seen in eq 5. Protein−ligand binding is entropically unfavorable in most cases due to loss of translational, rotational, and conformational degrees of freedom.[38,53−55] Several scoring functions estimate the conformational entropy component of the free energy change via the number of rotatable bonds in the ligand or the protein and ligand together.[17,25,26] This measure provides a good estimate of the degrees of freedom lost by both the protein and ligand on binding.

Another way to calculate the conformational entropy component is to calculate the vibrational frequencies of the ligand in complex[37] with the entire QM region, as well as the protein and ligand alone. By capturing these effects, the change in translational, rotational, and vibrational degrees of freedom upon binding can be estimated, and the entropic effect on the ligand in the protein field or on the ligand and the protein side chains in the active site that interact with the ligand can be determined. This gives a more accurate estimate of the entropic penalty at the cost of computing the vibrational frequencies. In this implementation, after QM/MM conjugate gradient optimization, the frequencies of the optimized ligand and protein alone and in complex with the protein were calculated at the AM1 level using DivCon and a partial Hessian vibrational analysis (PHVA).[3] This allows the frequencies of only the minimized region, either the ligand alone in the protein field or the entire QM region, to be calculated, excluding the rest of the system, and has been found to accurately reproduce the appropriate frequencies.[3,56]

Many of the systems being considered here are fairly large overall, and therefore, the minimization steps will not fully minimize all gradients to a near-zero value. This is due partially to the minimization scheme used and partially to our interest in keeping the binding affinity prediction quick enough for large-scale applications. Thus, a small number of imaginary frequencies may appear from the diagonalization of the Hessian matrix, especially when calculating the frequencies of the entire QM region, comprised of the ligand and protein side chains. In these cases, any imaginary vibrational frequencies found were not included in the calculation of the vibrational entropy. Some low-frequency vibrational modes may be disregarded, but this calculation can still provide an estimate of the vibrational entropy change due to binding.

From these frequencies the vibrational entropy, energy, and zero-point energy can be calculated from the normal-mode frequencies. The vibrational entropy component accounts for the change in entropy due to the gain of six vibrational degrees of freedom and loss of translational and rotational degrees of freedom when the ligand binds to the protein. The vibrational energy component represents the internal thermal energy change from molecular vibrations upon ligand binding. Here, the vibrational entropy is calculated either by finding the vibrations of the ligand in complex and free, which has been shown to be a good approximation of the degrees of freedom of the protein and ligand system,[53] or by calculating the frequencies of the QM/MM components including the entire QM region's side chains. In this study both of these methods are explored in the interest of measuring accuracy vs cost. The zero-point energy (ZPE) corrects the energy of the system up from the bottom of the harmonic oscillator well to the lowest vibrational quantum level, accounting for vibrations occurring even at 0 K. In this study, the vibrational energy and ZPE are already implicitly included in the calculation due to the parametric way in which the semiempirical methods are developed (fit against experimental heats of formation) and therefore are excluded from the scoring function to avoid double counting these properties. Calculating these vibrational frequencies provides the estimate to the entropy change in the gas phase that we use in our overall scoring function.

The final part of the binding free energy we consider is the solvation free energy change of the system. We calculate this as the sum of enthalpy and entropy terms:

$$\Delta G_{\text{solv}} = \Delta H_{\text{solv}} - T\Delta S_{\text{SASA}} \quad (8)$$

It has been shown that an implicit model of solvation appropriately describes the solvent interactions that occur in protein−ligand binding to account for the solvation effects.[38,57] This term can roughly be described as the free energy associated with the desolvation of the active site and ligand upon ligand binding.

In this study an implicit solvent Poisson−Boltzmann self-consistent reaction field (PB/SCRF) solvation model was used to calculate the solvation enthalpy.[58] This method essentially calculates the energy of the solute with and

without the presence of solvent, allowing only the QM region to polarize when the solvent is added to the calculation. This gives an approximate solvation enthalpy for the active site, allowing the charges in that area to fluctuate in the solvent field, while holding the remaining charges in the system fixed. A previous study by Merz et al. found an unsigned average error of 19.2 kcal/mol for four small proteins using the same QM region cutoff size. Using a Poisson−Boltzmann method also means that an internal and external dielectric must be set to properly capture the polarization of the QM region.[59] In this study CM1[60] charges, an external dielectric constant of 80, and an internal dielectric constant of 1.0 were used. The dielectric constant is used to estimate the polarizability of the region it is assigned to. In fixed charge methods, values of 3.0 and 4.0 are commonly used for an estimate of polarizability. However, here a value of 1.0 was used because the QM region is permitted to polarize in the solvent field due to the presence of electronic degrees of freedom. Using this QM/MM PB method allowed the solvation enthalpy of the protein, ligand, and complex to be determined, yielding the overall enthalpic cost of desolvating the ligand and protein active site.

There is also an entropic solvation term to be considered on binding due to the displacement of water molecules from the active site, which plays an important role in binding.[38,61] Changes in solvent-exposed surface area upon ligand binding have a correlation to solvent entropy,[62] and some scoring functions successfully use the term as a measure of solvent entropy.[11,37,38,63] In this study the solvent-accessible surface area of the heavy atoms, regardless of the polarity of the atoms, was used to estimate the solvation entropy of the binding process. This was done by running a 1.4 Å probe over the surface of the protein, ligand, and complex, which yielded the SASA[64] of each respective piece of the protein−ligand binding calculation. This surface area difference gives an estimate of the solvent entropy gained upon complexation based on the parameters from Legrand and Merz.[64] Combining the enthalpy term from the PB equations and the entropy term of the SASA approximation, the free energy change on solvation can be calculated using eq 8. The enthalpy term is simply the difference in enthalpy between the solvated and unsolvated protein, ligand, or complex, the solvation energy, and the entropy term is approximated using the SASA term.

After calculation of all of these individual components, a scoring function can be constructed to calculate the binding affinity:

$$\Delta G_{\text{bind}} = \Delta G_{\text{gas}}^{\text{bind}} + \Delta E_{\text{PB}} \tag{9}$$

In this equation, $\Delta G_{\text{gas}}^{\text{bind}}$ is the energy change from the ESCF of the unbound protein and ligand going to the complex containing the $\Delta S_{\text{vib}}$ and $\Delta S_{\text{SASA}}$ entropies and $\Delta E_{\text{PB}}$ is the Poisson−Boltzmann energy change. Combining all of these individual terms provided an overall equation that allowed the binding free energy to be calculated from individual energy components.

**Regression Analysis.** The method described here was assessed by comparing its predictions against experimental data using both the square of the sample correlation coefficient, $R^2$, and the standard deviation, SD:

$$R^2 = 1 - \frac{\sum_i (Y_i - f_i)^2}{\sum_i (Y_i - \bar{Y})^2} \tag{10}$$

$$\text{SD} = \sqrt{\frac{1}{N}\sum_i (Y_i - f_i)^2} \tag{11}$$

where $f_i$ are the predicted free energies of binding, $Y_i$ are the experimental free energies of binding, and $\bar{Y}$ is their average.

We also attempted to improve the method by using multiple linear regression (MLR) of the terms in the energy function to predict the binding free energies. MLR defines a relationship between a dependent variable and independent variables using a least-squares method. MLR produces a linear equation where $X_1$, $X_2$, ... are independent variables, here components of the scoring function, and $Y$ is the dependent variable, the binding affinity:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \varepsilon_i \tag{12}$$

for data points $i = 1, 2, ..., N$. By using these methods, not only can the predictive ability of a scoring function be estimated, but weights for individual terms ($\beta_0$, $\beta_1$, ...) in the scoring function can be determined to give them more influence in the overall score and improve the scoring function.[9,25] $R^2$ and the standard deviation can then be calculated from the values predicted by the linear regression

$$f_i = \beta_0 + \beta_1 \Delta E_{\text{SCF}} - \beta_2 T \Delta S_{\text{vib}} + \beta_3 \Delta E_{\text{PB}} + \beta_4 \Delta S_{\text{SASA}} \tag{13}$$

in comparison with the experimental data, using eqs 10 and 11.

## Results and Discussion

The proteins, ligands, and complexes used in this study were taken from a full QM study undertaken by Raha and Merz,[25] consisting of 18 carbonic anhydrase (CA) and 5 carboxypeptidase (CPA) complexes with known experimental binding free energies and resolutions better than 2.5 Å. Table 1 summarizes the proteins and their ligands, resolution, and experimental binding affinities as well as the number of QM atoms for each. One of DivCon's features is its linear-scaling capabilities, which allows larger than usual systems to be used in semiempirical calculations.[22,23] The DivCon method splits a protein into separate groups of atoms, in this case amino acids. These groups include a buffer region around them to account for polarization effects on the central group. The SCFs of these individual groups and their buffers are then calculated separately. This is faster than performing an SCF on the entire protein at once, since it is quicker to do many small diagonalizations than one very large one for each SCF cycle. This allows the method to be useful for large systems. Here, the crossover point for standard versus divide and conquer calculations was determined to be approximately

***Table 1.*** PDB ID, Resolution (Å), Inhibitor, Type, and Experimental $\Delta G$ of the 18 Carbonic Anhydrase (CA) Complexes and 5 Carboxypeptidase (CPA) Complexes Used in This Study[a]

| PDB ID | resolution | inhibitor | type | $\Delta G$(exp) | protein | ligand | complex |
|---|---|---|---|---|---|---|---|
| 1a42 | 2.25 | brinzolamide | CA | −13.66 | 210 | 44 | 250 |
| 1am6 | 2.1 | methyl hydroxamate | CA | −5.98 | 190 | 10 | 196 |
| 1bcd | 1.9 | methyl sulfonamide | CA | −5.39 | 210 | 10 | 216 |
| 1bn1 | 2.1 | AL5917 | CA | −12.90 | 210 | 35 | 241 |
| 1bn3 | 2.2 | AL6528 | CA | −13.66 | 210 | 35 | 241 |
| 1bn4 | 2.1 | AL5927 | CA | −12.86 | 210 | 36 | 242 |
| 1bnn | 2.3 | AL7183 | CA | −13.82 | 210 | 37 | 243 |
| 1bnq | 2.4 | AL4623 | CA | −13.11 | 210 | 41 | 247 |
| 1bnt | 2.15 | AL5424 | CA | −13.54 | 210 | 38 | 244 |
| 1bnu | 2.15 | AL5300 | CA | −13.40 | 210 | 34 | 240 |
| 1bnv | 2.4 | AL7099 | CA | −12.12 | 210 | 42 | 248 |
| 1bnw | 2.25 | AL5415 | CA | −12.54 | 210 | 29 | 235 |
| 1cil | 1.6 | ETS | CA | −12.90 | 160 | 25 | 182 |
| 1cim | 2.1 | PTS | CA | −12.19 | 210 | 35 | 241 |
| 1cin | 2.1 | MTS | CA | −12.06 | 210 | 29 | 235 |
| 1cnw | 2.0 | EG1 | CA | −10.67 | 210 | 33 | 239 |
| 1cnx | 1.9 | EG2 | CA | −10.12 | 264 | 51 | 311 |
| 1cny | 2.3 | EG3 | CA | −10.85 | 210 | 44 | 250 |
| 1cbx | 1.54 | L-benzylsuccinate | CPA | −8.77 | 210 | 64 | 270 |
| 3cpa | 1.54 | GY | CPA | −5.37 | 123 | 31 | 150 |
| 6cpa | 1.54 | ZAAP(O)F | CPA | −15.93 | 160 | 57 | 214 |
| 7cpa | 2.0 | BZ-FVP(O)F | CPA | −19.30 | 141 | 73 | 211 |
| 8cpa | 1.54 | BZ-AGP(O)f | CPA | −12.66 | 123 | 54 | 174 |

[a] The experimental binding free energy for each complex was calculated from the $K_i$ as $-RT \ln(K_i)$, giving the binding free energy in kilocalories per mole. The last three columns show the number of QM atoms in each system for each metalloenzyme.

600 atoms in the QM region. We therefore determined QM regions with this limitation in mind.

In the following, several different possible variables and scoring function components will be explored and compared to experimental binding free energies for these metalloenzymes to assess this function's viability. This procedure will produce the best components for use in the scoring function while also providing details on important considerations for the scoring function. We will also examine the use of MLR weights to determine the influence this statistical method will have on the predictions. Finally, an analysis of different methods to include the vibrational entropy will be undertaken to explore ways to enhance the computational performance.

**ESCF Is More Predictive Than Total Energy.** In the case of a QM/MM calculation an important consideration is whether to use the total energy of the system or only the QM region's energy. The QM region's energy, ESCF, encompasses only the energy of the QM region, including all the changes in the electronic terms associated with binding, while the total energy encompasses the ESCF energy as well as the MM energy terms, such as bond and angle terms. Either of these terms may be appropriate to describe protein−ligand binding, and their effect on binding affinity prediction must be examined to determine the best one to use.

In this case, we found that the ESCF energy is a marked improvement over the total energy. As calculated with eqs 10 and 11, the square of the correlation coefficient, $R^2$, for the total energy is 0.56 with a standard deviation of 2.09 kcal/mol, while $R^2$ for the ESCF energy is 0.64 with a standard deviation of 1.89 kcal/mol. These binding affinities were calculated after two minimization runs with a nonbond cutoff of 100 Å. It can be argued that only the ESCF energy is essential in a QM/MM scoring function because the most detail is focused on the area of greatest interest. Since the active site and ligand, when present, are defined as the QM region, it is clear that the greatest energy changes are located within the QM region and that to a first approximation the MM region is largely unaffected. Indeed, including the total energy of the protein may have no effect, or even lower the predictive ability of a scoring function as random movements unassociated with binding may occur in the MM region, introducing spurious energy terms to the total energy. These terms may especially be found at the periphery of the protein, which will have the least effect on protein−ligand binding assuming nonallosteric interactions. These distant changes may reduce the predictive ability through long-range interactions that do not reflect binding per se, but are an artifact of our computational procedure. For this reason we consider it prudent to disregard the total energy of the protein in favor of using just the ESCF energy of the smaller QM region.

**A Long-Range Cutoff of 100 Å Behaves Better Than a Cutoff of 10 Å.** Another important consideration is the effect of the long-range nonbond cutoff. Atoms throughout the protein may affect the electronic properties of the QM active site and ligand, which may in turn enhance or diminish protein−ligand binding. It is important to account for this when calculating the binding free energy of protein−ligand complexes. The binding affinity of the protein set was calculated using the scoring function with two steps of minimization, the ESCF energy and a cutoff of 10 and 100 Å, to explore this parameter's effects on the calculated binding affinity.

The long-range cutoff has a large impact on the predictive ability of the scoring function. Using a cutoff, for both the QM and MM regions, of 10 Å gives an $R^2$ of 0.36 with a standard deviation of 2.51 kcal/mol. However, using the larger cutoff of 100 Å yields an $R^2$ of 0.64 with a standard deviation of 1.88 kcal/mol. For these relatively small systems this cutoff includes the entire protein in the nonbond cutoff

calculation. The importance of long-range cutoffs is well-known and, not surprisingly, appears to be no less important in these binding affinity calculations than in molecular dynamics simulations.[49]

**A Single Minimization Cycle and the ESCF Scoring Function Perform Well.** In these calculations, the QM/MM energy of the protein, the ligand, and the protein−ligand complex are minimized using the steepest decent and conjugate gradient minimizers within AMBER. The starting geometries had the hydrogens minimized while the heavy atoms were held fixed, but the rest of the protein structure is that of the crystal structure found in the PDB. Minimizing the protein allows the QM and MM regions to relax in their molecular environment, lowering the overall energy of the protein and providing a good starting point for further calculations. These QM/MM minimizations also relax the ligand in both the free and bound states, giving insight into possible structural changes of the ligand. Minimization also makes it possible to properly calculate the vibrational frequencies.

The predictive ability of the scoring function on unminimized structures was also examined. In these studies the heavy atoms were obtained directly from the PDB, while the protons were added with LEAP and allowed to minimize. From these structures the vacuum interaction energy, solvation free enthalpy, and solvation entropy were calculated. For these calculations the vibrational entropy component was estimated using the number of rotational bonds in the ligand, assigning a 1 kcal/mol penalty to each bond as used by Raha et al.[25,26] The number of rotatable bonds is used here because using the vibrational frequencies without minimization results in large numbers of imaginary frequencies and will not provide an accurate representation of the actual entropy change.

Using the best parameters found above, namely, the ESCF energy and a long-range cutoff of 100 Å, the predictive ability of the scoring function was tested as a function of how many minimizations were done, each minimization comprising a 500-step steepest descent followed by a maximum 1500-step conjugate gradient minimization. The number of minimization steps used has an interesting effect on the predictive ability of the scoring function. If the total energy of the protein is used, one minimization results in an $R^2$ of 0.48 with a standard deviation of 2.27 kcal/mol. Performing two minimizations and using the total energy of the system increases the squared correlation coefficient to 0.56 with a standard deviation of 2.09 kcal/mol. Using the ESCF energy, the correlation increases to 0.59 for one minimization and to 0.64 for two minimizations, while the standard deviations are 2.01 and 1.88 kcal/mol, respectively. The results of these predictions using structures without minimization and the vibrational entropy estimated by the number of rotatable bonds gives a correlation of 0.57 with a standard deviation of 2.07 kcal/mol. Figure 3 summarizes the results of each of these minimization runs, demonstrating the differences between ESCF and total energy for each minimization scheme.

It is interesting to note that, for one minimization step, the ESCF correlation coefficient exceeds the total energy

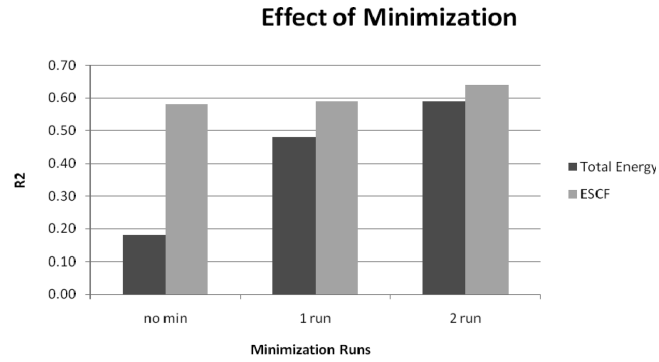


***Figure 3.*** Effect of various minimization protocols on binding affinity prediction.

coefficient, but the two binding affinity correlations are quite close for two minimization steps. This suggests that using the ESCF energy is an adequate measure of the heat of interaction and that a single minimization may be sufficient to predict binding affinities to save time. Using one minimization cycle would obviously save time over performing two cycles, so the ability to accurately calculate binding affinity at one minimization cycle is an important consideration. Figure 3 demonstrates that using the total energy of the system at one cycle is not as predictive as the ESCF energy while there is minimal difference between one and two cycles for the ESCF energy. It is also interesting to note that scoring using ESCF without minimization also produces a result relatively close to the minimized correlation. This observation is important because if a large screening effort were undertaken, this could be used to further reduce the cost while only slightly lowering the predictive quality over partial or full minimization.

**Binding Affinity Predictions Can Be Improved by MLR.** The binding affinities of the 23 ligands were recalculated using multiple linear regression (eqs 12 and 13), calculating weights for the four energy terms and the constant seen in eq 13. For these calculations, the components of the scoring function were calculated using the best parameters for each term as described above. The ESCF energy was used for the heat of interaction, a long-range cutoff of 100 Å was used, and, for thoroughness, two minimization cycles were performed even though one was shown to be adequate for the ESCF energy. For these binding affinity predictions, CM1 charges were used in the SCRF solvation calculation. Figure 4 demonstrates the results of both the simple sum and the fitted function.

These results show promise for using the QM/MM method in binding affinity prediction. Calculating the binding affinities of these zinc metalloproteins with MLR yields an $R^2$ of 0.71 with a standard deviation of 1.69 kcal/mol, compared to the results without any fitting ($R^2$ of 0.64, SD = 1.88).

**Using the Number of Rotatable Bonds To Estimate the Conformational Entropy Change.** As mentioned above, it is fairly common to use the number of rotatable bonds to estimate the vibrational entropy penalty. The use of this estimate was also examined in this study by counting the number of rotatable bonds on the ligand using the Autotors tool from AutoDock.[65] After the number of rotatable bonds

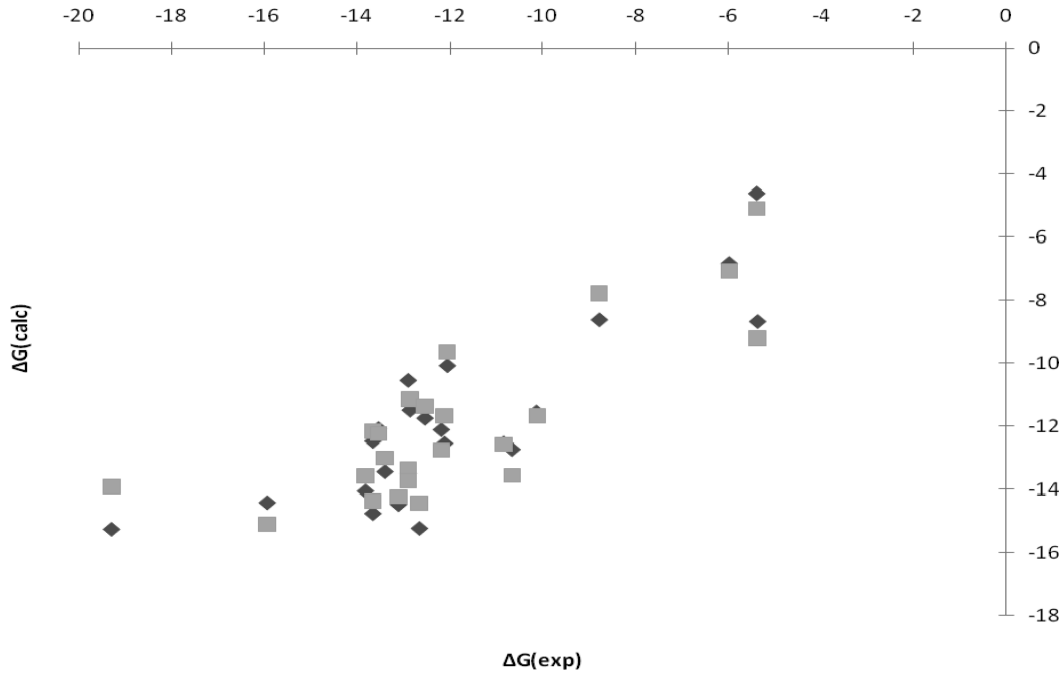


***Figure 4.*** Calculated vs experimental $\Delta G$ of binding for the 23 zinc protein−ligand complexes. Squares represent the binding affinity calculated without any fitting to experimental data ($R^2 = 0.64$). Tilted squares represent the calculated binding affinity after fitting components of the scoring function using MLR ($R^2 = 0.71$). All units are kilocalories per mole.

in the ligand was counted, it was assumed that each of these bonds would be held fixed upon binding and each bond was responsible for 1 kcal/mol, representing the conformational degrees of freedom lost on complexation. Using energies and the surface area from the minimized structures, these calculations resulted in a correlation coefficient of 0.63 and a standard deviation of 1.92 kcal/mol without fitting. Without the minimization, these values are 0.57 and 2.07 kcal/mol (Figure 3). The correlation increases to 0.70 and the standard deviation decreases to 1.72 kcal/mol with MLR fitting. Again, the results without fitting are close to the results obtained by calculating the frequencies of the entire QM region, while the results with fitting are slightly worse than those calculated using only the ligand frequency analysis. This is an interesting result, confirming that using the number of rotatable bonds in the ligand as an estimate of the conformational entropy change may be sufficient to capture its effect on the predictive ability of this scoring function. The rotatable bond method may be used instead of a frequency calculation to save time if, for example, the ligand or data set is particularly large. It is also interesting to note that, at least for this set of 23 ligands, the number of rotatable bonds correlates with the vibrational entropy of the ligand calculated by DivCon with a correlation coefficient of 0.81, as seen in Figure 5. This presents more evidence that using the number of rotatable bonds is an acceptable estimate for the vibrational entropy of the ligand, and this can be capitalized upon to reduce the time from several hours for a full QM region vibrational calculation to a few seconds for the number of rotatable bonds if desired. Moreover, via more advanced analyses of free and bound ligands, it may be possible to create a rotatable bond model that tracks high-level computed results with an even greater accuracy.

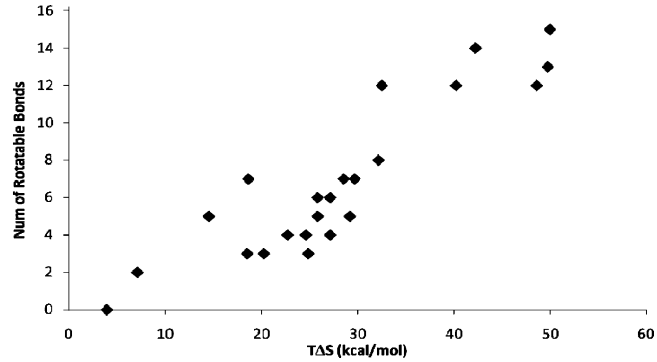


***Figure 5.*** Correlation of $T\Delta S_{vib}$ with the number of rotatable bonds for each ligand in the set from ligand-only frequency calculations ($R^2 = 0.81$).

**Using the Vibrational Frequencies of the Ligand Alone To Estimate the Conformational Entropy Change.** As a way to accelerate scoring calculations, some binding affinity calculations ignore or simply estimate the vibrational entropy component. It has been found that a good estimate can be achieved using the number of rotatable bonds in the ligand[4,8] or the change in the number of freely rotatable bonds in the ligand and protein based on solvent exposure.[25,26,66] These two methods used to estimate vibrational entropy are extremely quick, as the number of rotatable bonds is easily calculated, but are not always as accurate as desired and do not necessarily reflect the properties of the ligand. In this situation a compromise must again be made between accuracy and computation time to decide whether this estimate is accurate enough.

To account for the entropic term in this scoring function, the vibrational frequencies of some or all of the QM atoms can be calculated. This provides a more physical representation of what is happening when the ligand binds than a simple

count of the rotatable bonds. It also provides the additional advantage of calculating the frequencies in the field of the entire protein, allowing short- and long-range interactions to influence the results. As mentioned above, due to limitations in the minimization scheme used here, imaginary frequencies were often found from the normal-mode analysis. With the minimizers used in AMBER we were unable to eliminate all imaginary frequencies. Advanced minimizers or very long minimization runs might be able to mitigate this issue and may be investigated in future efforts. These imaginary frequencies were more prevalent for larger QM regions, but still comprised only a very small portion of the overall frequencies calculated, and all were subsequently left out of the total vibrational entropy calculations. Excluding these frequencies may leave out some vibrational contributions, but overall the vibrational entropy calculated here can still provide a reasonable estimate of the entropy change associated with binding.

In the previous sections we used the vibrational entropy calculated from the entire QM region, except when no minimization was done. These frequencies capture effects in both the ligand and the protein's active site where all of the major vibrational changes will occur from binding. While in principle the most accurate, this method is also the most expensive because the active site and ligand must be fully minimized and the Hessian computation is time-consuming. This method may work for small proteins or small QM regions, but may prove to be much more difficult and time-consuming for larger proteins or QM regions; these frequency calculations are, however, fairly trivial to parallelize. It is for these reasons that in this section we will examine various ways of calculating this vibrational entropy component and the impact they have on the scoring function's predictive ability.

An alternative afforded by the QM/MM method is to calculate the vibrational frequencies of just the ligand in the protein field, as successfully applied by Grater et al.[37] In this calculation the entire protein is frozen and the vibrational frequencies of the ligand, in the electronic field of the protein, are calculated. These frequencies are then compared to those for the ligand alone to find the change in vibrational entropy. Since the ligand will be a more manageable size than the entire QM region, this makes the minimization requirement easier to meet, as well as reducing the number of atoms in the normal-mode calculation. These predictions were made with the same criteria as above, using the two-step minimization and the ESCF energy of the QM region while varying the manner in which the frequencies were calculated. In the case of these 23 zinc metalloenzyme complexes, using the frequency of just the ligand results in an $R^2$ of 0.63 with a standard deviation of 1.89 kcal/mol without fitting. This can be compared to the results of the PHVA of the entire QM region, which yielded results of $R^2 = 0.64$ with a standard deviation of 1.88 kcal/mol. With MLR fitting, the correlation coefficient of the ligand-only vibrations increases to 0.72 with a standard deviation of 1.66 kcal/mol, while the frequencies of the full QM region yield a correlation of 0.71 and a standard deviation of 1.69 kcal/mol.

***Table 2.*** Time (min) Required To Calculate the QM Region and the Ligand-Only Vibrational Frequencies with the QM/MM Method for the Protein−Ligand Complex

| PDB ID | QM region | ligand-only |
|---|---|---|
| 1a42 | 868.54 | 8.94 |
| 1am6 | 416.45 | 0.37 |
| 1bcd | 580.17 | 0.44 |
| 1bn1 | 804.06 | 29.05 |
| 1bn3 | 802.76 | 6.50 |
| 1bn4 | 794.45 | 6.74 |
| 1bnn | 818.37 | 8.38 |
| 1bnq | 825.55 | 8.09 |
| 1bnt | 846.12 | 7.30 |
| 1bnu | 785.09 | 10.03 |
| 1bnv | 857.38 | 10.12 |
| 1bnw | 735.04 | 5.76 |
| 1cbx | 378.02 | 3.43 |
| 1cil | 797.64 | 5.57 |
| 1cim | 721.65 | 6.68 |
| 1cin | 759.12 | 5.96 |
| 1cnw | 1614.17 | 18.85 |
| 1cnx | 866.89 | 13.80 |
| 1cny | 1074.23 | 34.21 |
| 3cpa | 199.51 | 4.97 |
| 6cpa | 576.84 | 18.70 |
| 7cpa | 555.22 | 33.83 |
| 8cpa | 302.63 | 16.84 |
| average | 738.26 | 11.50 |

The unfitted and fitted results demonstrate that the extra information obtained from a full QM PHVA does not favorably impact our ability to predict protein−ligand binding. This result is somewhat unexpected because the full QM PHVA quantifies entropy changes in the protein side chains, but according to these results, these entropy changes do not have a large impact on the predictive ability of the scoring function for this set. Again, this impact may increase as the size of the QM region increases and more side chains are accounted for. However, these results must be viewed in terms of a cost−benefit analysis because calculating the vibrational frequencies can be quite expensive, so it is encouraging that the ligand-only results provide good predictions. Table 2 shows the time needed to calculate only the vibrational frequencies of the protein−ligand complex using the entire QM region and the ligand-only approach.

The range of vibrational entropies calculated here closely matches that from Schwarzl et al.[38] with a minimum of −9.85 kcal/mol and a maximum of 1.44 kcal/mol. The ligands in this study contain more rotatable bonds than those used in the study by Schwarzl, accounting for the larger range of entropies found here. Including the protein's vibrational entropy change moves the range of values to a minimum of −13.91 kcal/mol to a maximum of 4.91 kcal/mol. The appearance of this penalty in these vibrational calculations indicates a penalty to binding incurred due to loss of vibrational degrees of freedom in protein side chains. This is similar to the degrees of freedom lost in the ligand, but cannot be fully recovered with the ligand's new vibrational degrees of freedom, often leading to an overall binding penalty depending on the size of the ligand and composition of the side chains. It is also worth noting that the calculation time listed is that necessary for just one component of the vibrational entropy computation. When the protein's QM

***Table 3.*** Average Relative Timings for the Three Main Methods Used in the Study[a]

| method | relative time | $R^2$ |
|---|---|---|
| two minimization runs + QM vibrations | 8.83 | 0.64 |
| two minimization runs + ligand vibrations | 4.85 | 0.63 |
| no minimizations + rotatable bonds | 1.00 | 0.57 |

[a] Times are relative to the no minimizations and rotatable bond entropy estimate.

region is included in the vibrational analysis, an additional calculation must be undertaken to calculate the frequencies of the protein without the ligand and the ligand alone, which are additional steps that are time-consuming. These proteins are also relatively small, so the time would only increase even more with larger QM regions. This is an important consideration in the use of this scoring function, which could easily be modified to include the entire QM region frequencies or those of the ligand alone depending on the computational resources available and the amount of detail required. Overall, this analysis indicates that using the vibrations of the ligand alone may be enough to produce a good correlation, but more accuracy can potentially be gained using the frequencies of the entire QM region.

**Relative Timings.** To investigate the timing differences between the methods studied above, relative timings were compiled on the basis of average times for the various components. These will illustrate the timing differences between the methods in a manner that does not depend on the current level of technology, but gives an idea of the utility of these methods compared to each other. The relative timings for the QM/MM methods can be seen in Table 3 along with the correlation found using each method.

From Table 3 we can see that the full QM vibration calculation method is very expensive compared to the other two methods. Even the ligand-only vibration is expensive due to the slow minimization steps. This gap could perhaps be shortened by a quicker minimizer or perhaps using only one minimization run instead of two, which would be more acceptable for the ligand-only vibration calculations than the full QM vibrations due to negative frequencies. As the QM portion grows even larger, it could be expected that this scaling would become even worse for the minimization runs, necessitating a more efficient minimizer or skipping the minimization runs altogether and using the rotatable bond entropy estimate. The difference in the correlations is not great between all of these methods, but the timing is, so an appropriate method should be chosen depending on the number and size of the systems to be studied.

**Comparison of QM/MM with QM.** It is useful to compare the predictive abilities of this QM/MM study with those of the full QM study performed by Raha and Merz.[25] Both of these were done using the AM1 semiempirical Hamiltonian on the same set of metalloenzymes. This comparison will allow us to assess the QM/MM method's predictive ability in comparison to the full QM method and further determine the viability of the QM/MM approach for scoring. Raha and Merz reported the results without fitting and by fitting the solvation entropy term in the scoring function and calculated a correlation of 0.69 without fitting, while they report a higher correlation of 0.80 when using MLR on only the surface area terms of the organic heavy atoms in the protein. The QM/MM method here obtains a correlation of 0.71 using MLR fitting for all the terms, which is fairly close to the full QM prediction, while the predictions without fitting yield a correlation of 0.64. Not surprisingly, the QM/MM method does not do quite as well as the full QM method, but it is encouraging that it is qualitatively competitive. The correlations and standard deviations of these two studies are compiled in Table 4.

***Table 4.*** Comparison of Squared Correlation Coefficients ($R^2$) and Standard Deviations (SDs) between Full QM Results from Raha[25] and the QM/MM Method Described Here

| method | $R^2$ | SD (kcal/mol) |
|---|---|---|
| full QM, no fitting | 0.69 | 1.50 |
| full QM, MLR fitting | 0.80 | 1.18 |
| QM/MM, no fitting | 0.64 | 1.88 |
| QM/MM, MLR fitting | 0.71 | 1.69 |

**Charge Analysis.** An advantage of a full quantum or QM/MM method for these binding affinity calculations is that a QM method will be able to capture polarization and charge transfer (CT) effects. Polarization is generally considered an intramolecular term, representing the internal rearrangement of electrons, whereas charge transfer is an intermolecular term, which is an external source's electronic effects on the protein. CT has been shown to play an important role in binding,[44,45] and it is interesting to determine the degree of CT that occurs upon binding. When using a classical method, this term would generally be missed or estimated with a parametrized term, but a QM method allows the CT and polarization to have a more physical effect on the binding of the ligand. These CT effects can be especially relevant in a metalloenzyme complex due to the nature of metal ions and their bonds. Table 5 presents the charge transfer that occurs to the ligand in solution, when binding to the protein.

The table shows the charge transfer to the ligand involved in the binding process. CT was shown to contribute significantly to the interaction energy when solvating a protein, and it can safely be assumed that this is also an important factor in protein−ligand binding. Most of the carbonic anhydrase inhibitors gain approximately 0.9 electron upon binding when considering the CM1 charges, with the exception being 1am6, which only gains 0.71 electron. For the carboxypeptidase proteins the ligands give up some of their charge to the protein in all cases but 3cpa. This is due to the nature of the ligands: the ligand for 3cpa has a neutral charge, while the rest of the compounds have a negative charge. A standard molecular mechanics model would not be able to properly represent these, while a higher level QM method would better capture the effects, but might become too costly. This semiempirical QM/MM method provides a compromise for capturing CT effects, allowing them to be included while keeping the overall cost reasonable.

## Conclusions

We have presented a QM/MM method to calculate the binding free energy of protein−ligand complexes. This

***Table 5.*** Charge Transfer from the Protein to the Ligand (electrons) for Mulliken and CM1 Charges in Solution

| protein type | PDB ID | Mulliken | CM1 |
|---|---|---|---|
| CA | 1a42 | −0.57 | −0.87 |
| CA | 1am6 | −0.60 | −0.72 |
| CA | 1bcd | −0.64 | −0.91 |
| CA | 1bn1 | −0.68 | −0.84 |
| CA | 1bn3 | −0.58 | −0.88 |
| CA | 1bn4 | −0.58 | −0.88 |
| CA | 1bnn | −0.58 | −0.88 |
| CA | 1bnq | −0.57 | −0.87 |
| CA | 1bnt | −0.58 | −0.88 |
| CA | 1bnu | −0.58 | −0.88 |
| CA | 1bnv | −0.57 | −0.87 |
| CA | 1bnw | −0.58 | −0.87 |
| CA | 1cil | −0.57 | −0.87 |
| CA | 1cim | −0.57 | −0.87 |
| CA | 1cin | −0.61 | −0.90 |
| CA | 1cnw | −0.58 | −0.88 |
| CA | 1cnx | −0.58 | −0.88 |
| CA | 1cny | −0.58 | −0.88 |
| CPA | 1cbx | 0.37 | 0.35 |
| CPA | 3cpa | −0.56 | −0.61 |
| CPA | 6cpa | 0.43 | 0.41 |
| CPA | 7cpa | 0.49 | 0.46 |
| CPA | 8cpa | 0.27 | 0.26 |

method takes advantage of the linear scaling capabilities of DivCon to include a large number of atoms near the ligand in the semiempirical QM region. This allows electronic effects, which would be missed in a classical calculation, to be properly represented while keeping the computational cost low enough to be performed on a large library of protein−ligand complexes. This approach was successful at predicting the binding free energies of a set of 23 zinc metalloenzyme complexes. The scoring function performs well without any fitting, but through multiple linear regression, the function can be fit to experimental data and the predictive performance increases from a squared correlation coefficient of 0.64 to one of 0.71. This may only be the case within a single protein family, and an overall set of weights for general use may be necessary as in empirical scoring functions, but it shows that the method has potential.

This method may also take into account the vibrational entropy change of the ligand upon binding. The QM/MM method allows a unique perspective on this in that a normal-mode analysis can be conducted in the field of the protein's charges while charges further from the ligand remain fixed. This gives a more accurate representation of the vibrational modes available to the ligand, and therefore a more accurate representation of the vibrational entropy contribution to the overall binding affinity of the complex, but is computationally intensive. A simple estimate of counting rotatable bonds was also examined to estimate entropy change as a way to save computation time.

The contributions of various parameters for the predicted binding affinity were also investigated including long-range cutoffs and the use of the total energy of the system versus the QM energy for the heat of interaction. These studies indicated that the long-range cutoff used makes a significant difference in the predicted binding affinity. It was also found that the use of the ESCF energy to calculate the heat of interaction was preferable to the use of the total energy of the system. Using the ESCF energy, the number of minimization steps does not make as much of an impact, whereas increased minimization cycles greatly improve the predictive ability of the function when the total energy is used.

Although the scoring function presented is relatively good at predicting the binding affinities of zinc metalloenzyme complexes, there is room for improvement. Depending on the acceptable costs, a larger QM region can easily be chosen. This will provide a second shell of residues to interact with the ligand in the QM region, giving a better idea of the electronic effects involved in binding. Not only will this present a better representation of the charge transfer to and from the ligand, it will allow a better picture of the polarization due to the protein environment. Sampling of the system through MD snapshots similar to the MM-PBSA method could also be used to potentially improve the predictions of this scoring function. These snapshots will provide a sampling of the protein, potentially increasing the predictive ability while increasing the cost of the calculations. Using a purely classical simulation, this might present problems because each ligand would need to be parametrized properly, which is quite costly. However, a QM/MM method would allow the parametrization step to be skipped, and a QM region could be formulated to make a sampling of the different configurations tractable. All of the tools for the QM/MM scoring method are easily applied to the snapshots generated by a simulation to emulate the MM-PBSA method. A QM/MM simulation could also be constructed to include only the ligand in the QM region, allowing the protein to be sampled while removing the parametrization needs of the ligand.

Overall, the QM/MM scoring method presents good predictive trends at a reasonable cost, but does present some areas for future improvement. In its current form this method might be more useful verifying ligand poses or as a drug refinement step, but could be made more affordable through parallelization techniques and modification of the parameters used.

**Acknowledgment.** K.M.M. thanks the NIH (Grant GM044974) and R.D. thanks the NIH (Grants R01 GM84453 (R.D.) and P30 CA006927 (Fox Chase Cancer Center)) for financial support of this work.

## References

(1) Lyne, P. D. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(2) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(3) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.

(4) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.

(5) Wang, R.; Lu, Y.; Wang, S. *J. Med. Chem.* **2003**, *46*, 2287–2303.

(6) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.

(7) Trott, O.; Olson, A. J. *J. Comput. Chem.* **2009**, *31*, 455–461.

(8) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

(9) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(10) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.

(11) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. *J. Med. Chem.* **2004**, *47*, 3032–3047.

(12) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040–4048.

(13) Steinbrecher, T.; Case, D. A.; Labahn, A. *J. Med. Chem.* **2006**, *49*, 1837–1844.

(14) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(15) Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786–3791.

(16) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6825–6830.

(17) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770–2780.

(18) Brenk, R.; Vetter, S. W.; Boyce, S. E.; Goodin, D. B.; Shoichet, B. K. *J. Mol. Biol.* **2006**, *357*, 1449–1470.

(19) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. *Biochemistry* **2005**, *44*, 12316–12328.

(20) Nemoto, T.; Fedorov, D. G.; Uebayasi, M.; Kanazawa, K.; Kitaura, K.; Komeiji, Y. *Comput. Biol. Chem.* **2005**, *29*, 434–439.

(21) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(22) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1996**, *104*, 6643–6649.

(23) Dixon, S. L.; Merz, K. M., Jr. *J. Chem. Phys.* **1997**, *107*, 879–893.

(24) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. *J. Chem. Phys.* **2004**, *120*, 1145–1148.

(25) Raha, K.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.

(26) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558–4575.

(27) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; WollaCott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. *Drug Discovery Today* **2007**, *12*, 725–731.

(28) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(29) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(30) Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904–911.

(31) Friesner, R. A. *Adv. Protein Chem.* **2005**, *72*, 79–104.

(32) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.

(33) Senn, H. M.; Thiel, W. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(34) Peters, M. B.; Raha, K.; Merz, K. M., Jr. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 370–379.

(35) Zhou, T.; Huang, D.; Caflisch, A. *J. Med. Chem.* **2008**, *51*, 4280–4288.

(36) Cho, A. E.; Rinaldo, D. *J. Comput. Chem.* **2009**, 2609–2616.

(37) Grater, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474–10483.

(38) Schwarzl, S. M.; Tschopp, T. B.; Smith, J. C.; Fischer, S. *J. Comput. Chem.* **2002**, *23*, 1143–1149.

(39) Gleeson, M. P.; Gleeson, D. *J. Chem. Inf. Model.* **2009**, *49*, 670–677.

(40) Fong, P.; McNamara, J. P.; Hillier, I. H.; Bryce, R. A. *J. Chem. Inf. Model.* **2009**, *49*, 913–924.

(41) Anisimov, V. M.; Bugaenko, V. L. *J. Comput. Chem.* **2008**, 784–798.

(42) Illingworth, C. J. R.; Morris, G. M.; Parkes, K. E. B.; Snell, C. R.; Reynolds, C. A. *J. Phys. Chem. A* **2008**, *112*, 12157–12163.

(43) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915–931.

(44) Garcia-Viloca, M.; Truhlar, D. G.; Gao, J. *J. Mol. Biol.* **2003**, *327*, 549–560.

(45) Ji, C. G.; Zhang, J. Z. *J. Am. Chem. Soc.* **2008**, *130*, 17129–17133.

(46) Illingworth, C. J. R.; Parkes, K. E.; Snell, C. R.; Marti, S.; Moliner, V.; Reynolds, C. A. *Mol. Phys.* **2008**, *106*, 1511–1515.

(47) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.

(48) Field, M. J.; Albe, M.; Bret, C.; Proust-De Martin, F.; Thomas, A. *J. Comput. Chem.* **2000**, *21*, 1088–1100.

(49) Walker, R. C.; Crowley, M. F.; Case, D. A. *J. Comput. Chem.* **2008**, *29*, 1019–1031.

(50) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.

(51) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(52) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(53) Fischer, S.; Smith, J. C.; Verma, C. S. *J. Phys. Chem. B* **2001**, *105*, 8050–8055.

(54) Chang, C. E. A.; Chen, W.; Gilson, M. K. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534–1539.

(55) Murray, C. W.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 741–753.

(56) Calvin, M. D.; Head, J. D.; Jin, S. Q. *Surf. Sci.* **1996**, *345*, 161–172.

(57) Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.

(58) Hayik, S. A.; Liao, N.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2008**, *4*, 1200–1207.

(59) Archontis, G.; Simonson, T.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 307–327.

(60) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87–110.

(61) Vajda, S.; Weng, Z. P.; Rosenfeld, R.; Delisi, C. *Biochemistry* **1994**, *33*, 13977–13988.

(62) Bardi, J. S.; Luque, I.; Freire, E. *Biochemistry* **1997**, *36*, 6588–6596.

(63) Velec, H. F.; Gohlke, H.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6296–6303.

(64) Legrand, S. M.; Merz, K. M., Jr. *J. Comput. Chem.* **1993**, *14*, 349–352.

(65) 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*, 1639–1662.

(66) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505–524.

CT100315G