LINEAR-SCALING METHODOLOGY IN LARGE-SCALE AB INITIO ELECTRONIC STRUCTURE CALCULATIONS AND APPLICATIONS IN BIOLOGICAL STUDIES

By

XIAO HE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2010

To my dear parents

ACKNOWLEDGMENTS

I would like to thank my advisor, Kennie for his generous support and guidance. I thank all present and past members of the Merz group for their help. Special thanks to my previous advisor, John Zhang and my colleagues: Laszlo, Bing, Ning, and Xue, for various helpful discussions. Lastly, I’m deeply thankful for the abundant love and support of my parents. Without them, this dissertation would not have been written.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HF	Hartree-Fock theory
DFT	Density Functional theory
MP2	Second-order Møller-Plesset perturbation theory
CC	Coupled-cluster theory
SCF	Self-consistent field
DFT	Density functional theory
LMO	Localized molecular orbitals
DC	Divide-and-conquer algorithm
MFCC	Molecular fractionation with conjugated caps method
SAD	Superposition of atomic densities
AF-QM/MM	Automated fragmentation quantum mechanics/molecular mechanics approach
FMO	Fragment molecular orbital approach
PCM	Polarizable continuum model
CBS	Complete basis set limit
RMSE	Root mean square error
MUE	Mean unsigned error
MSE	Mean signed error

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

LINEAR-SCALING METHODOLOGY IN LARGE-SCALE AB INITO ELECTRONIC STRUCTURE CALCULATIONS AND APPLICATIONS IN BIOLOGICAL STUDIES

By

Xiao He

May 2010

Chair: Kenneth M. Merz, Jr.
Major: Chemistry

The ability to perform *ab initio* electronic structure calculations with times that scale linearly with the system size is one of the central aims in theoretical chemistry. In this dissertation, the implementation of the divide-and-conquer (DC) algorithm, an algorithm with the potential to aid in linear scaling capability in Hartree-Fock (HF) and second-order Møller-Plesset perturbation (MP2) calculations, is discussed. Standard HF calculations solve the Roothaan-Hall equations for the whole system; in the DC-HF approach, the diagonalization of the Fock matrix is carried out on smaller subsystems. For DC-MP2 calculations, after localized molecular orbitals of each subsystem are obtained from the DC-HF calculations, the correlation energy of the whole system can be derived by taking the sum of the local electron correlation of each subsystem. Preliminary DC-MP2 results on extended polyglycine systems show the linear-scaling behavior.

We have also proposed an automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM) to routinely calculate *ab initio* protein NMR chemical shielding constants. The AF-QM/MM method is linear-scaling and trivially parallel. A general fragmentation scheme is employed to generate each

residue-centric region which is treated by quantum mechanics, and the environmental electrostatic field is described with molecular mechanics. The AF-QM/MM method shows good agreement with standard self-consistent field calculations of the NMR chemical shieldings for the mini-protein Trp-cage.

This dissertation also deals with an application of these faster implementations of *ab initio* methods to examine future uses of our code. Our linear-scaling approach is still in the development stages, we therefore chose to use the fastest currently available method for carrying out *ab initio* electronic structure calculations, the fragment-molecular-orbital (FMO) approach. By utilizing the available software GAMESS-US, we employed both FMO-HF and FMO-MP2 calculations in conjunction with the Polarizable Continuum Model on the native structures of two proteins and their corresponding computer-generated decoy sets. We show the sum of the HF energy and force field (LJ6) derived dispersion energy (HF + LJ6) is well correlated with the energies obtained using second-order MP2 theory. In one of the two examples studied the correlation energy as well as the empirical dispersive energy term was able to discriminate between native and decoy structures. On the other hand, for the second protein we studied, neither the correlation energy nor dispersion energy showed discriminative capabilities; however, the *ab initio* MP2 energy and the HF+LJ6 both ranked the native structure correctly.

CHAPTER 1
INTRODUCTION

Ab initio quantum mechanical methods have been developed over the past several decades and successfully applied to study the chemical properties form small to moderate-sized molecules. The routine application of these full quantum mechanical calculations on macromolecules (molecules containing greater than 500 atoms) continues to pose great challenges for theoretical chemists. The major limitation of *ab initio* methods is the scaling problem, since the computational cost of *ab initio* methods increases considerably as the size of the molecule increases. For instance, Hartree-Fock (HF)[1] and Density Functional Theory (DFT)[2] scale as O(N^4), MP2 scales as O(N^5) and the coupled cluster(CC)[3] method that includes single and double excitations (CCSD) scales as O(N^6). In modern HF calculations, the computational cost for the 2-electron integrals can be reduced from O(N^4) to O(N^2) using a simple screening technique[4]. Hence, the dominant step for large molecules becomes the matrix diagonalization, which scales as O(N^3). In this thesis, our goal was to reduce the computational cost of the diagonalization step in HF calculations to linear with system size.

The state-of-the-art linear-scaling algorithms which make the computational cost scale linearly $O(N)$ with the system size, have attracted the focus of many theorists during the past decade.[5-15] The aim of our current research is to further develop the divide-and-conquer (DC)[16-22] methodology to aid in the application of *ab initio* methods on the larger molecules (see Chapters 2 and 3). In the DC algorithm, the total system is divided into small fragments. Atoms within adjustable buffer regions surrounding each fragment are included in the calculations to preserve the chemical environment of the divided subsystem. A set of local Roothaan-Hall equations is then solved for each subsystem and an approximate full density matrix of the entire molecular system is built up from subsystem contributions. By solving the HF self-consistent field (SCF) equation iteratively, the final converged full density matrix is used to obtain the total energy of the entire system. In this manner, linear scaling of the Fock matrix diagonalization step is achieved as a result of the fact that a set of smaller subsystem Fock matrices is diagonalized in the DC-HF approach rather than the global Fock matrix diagonalization for traditional HF calculations. In the framework of DC, MP2 electron correlation energy for the entire system can be derived from local correlation of the localized molecular orbitals (LMOs) on each subsystem. By decomposing the total electron correlation energy into contributions from each subsystem, the correlation energy of the whole system is the sum of the subsystem-based correlation energies, so that the computational cost on MP2 electron correlation energy becomes linear-scaling as well. Furthermore, divide-and-conquer calculations may be efficiently parallelized on massive computer nodes because the individual subsystem calculations are solved separately. In the DC-HF implementation, the memory usage will be increased linearly as the size of the system increases. On the other hand, after DC-HF calculation is solved, for DC-MP2 electron correlation energy calculation, the memory requirement is independent of the size of the whole system because the electron correlation energy can be calculated for each subsystem separately.

Various applications on biological systems could be studied using linear-scaling *ab initio* approaches. Chapter 4 discusses an automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM)[23] to routinely calculate *ab initio* protein NMR chemical shielding constants. Chapter 5 examines a quantum mechanical(QM) energy-based "scoring" function[24] that can routinely discriminate natively folded proteins from the non-native conformations.[25] Based on the thermodynamic hypothesis, which states that the native conformation has the lowest free energy relative to misfolded states[26], current effort focuses on looking for reliable physics-based potentials that can distinguish native states from non-native ones.[27-31] Importantly, the free energy of a native three-dimensional structure is only 5-15 Kcal/mol less than misfolded states[32,33]; hence, it is clear that the final solution to this problem will require very high accuracy. We employed both FMO-HF and FMO-MP2 calculations in conjunction with the Polarizable Continuum Model (PCM) on the native structures of two proteins and their corresponding computer-generated decoy sets.

Accurate benchmark calculations of gas-phase basicities of small molecules are presented in Chapter 6 and compared with available experimental results.[34] The optimized geometries and thermochemical analyses were obtained from MP2/aug-cc-pVTZ calculations. Two different *ab initio* electron-correlated methods MP2 and CCSD(T) were employed for subsequent gas-phase basicity calculations and the single point energies were extrapolated to the complete basis set (CBS) limit. We have proposed an efficient approach to predict gas-phase basicities of small molecules within chemical accuracy.

CHAPTER 2
THEORY AND METHODS

2.1 The Hartree-Fock Approximation

The Hartree-Fock (HF) method uses a single determinant wave function and approximates the electron repulsion by a mean field potential. Based on Linear Combination of Atomic Orbitals (LCAO) approximation, Molecular Orbitals (MO) $(\psi_1, \psi_2, \cdots \psi_N)$ are expanded by a set of atomic orbitals (AO) $(\phi_1, \phi_2, \cdots \phi_N)$

$$\psi_i = \sum_{\mu=1}^{N} C_{\mu i} \phi_\mu \tag{2-1}$$

For closed-shell systems the density matrix is given by

$$P_{\mu\nu} = 2\sum_{i=1}^{n_{occ}} C_{\mu i} C_{\nu i}^{*} \tag{2-2}$$

In the Hartree-Fock self-consistent field (SCF) formalism, the MO coefficients $C_{\mu i}$ are determined by solving the Roothann-Hall equation self-consistently.[1]

$$FC = SCE \tag{2-3}$$

where S is the overlap matrix ($S_{\mu\nu} = \int dr \varphi_\mu^{*}(r) \varphi_\nu(r)$), C represents the MO coefficient matrix, E is the diagonal MO energy matrix, and F is the Fock matrix, which is defined by

$$F_{\mu\nu} = H_{\mu\nu}^{core} + \sum_{\lambda\delta} P_{\lambda\delta}[(\mu\nu \mid \delta\lambda) - \frac{1}{2}(\mu\lambda \mid \delta\nu)] \tag{2-4}$$

where $H_{\mu\nu}^{core}$ is the one-electron core-Hamiltonian Matrix

$$H_{\mu\nu}^{core} = \int dr \phi_\mu^{*}(r) \left[-\frac{1}{2}\nabla_r^2 - \sum_{A=1}^{N_{atoms}} \frac{Z_A}{|r - r_A|} \right] \phi_\nu(r) \tag{2-5}$$

and $(\mu\nu \mid \delta\lambda)$ is the two-electron integral in chemists' notation.

$$(\mu\nu\,|\,\delta\lambda) = \iint dr_1 dr_2 \phi_\mu^*(r_1)\phi_\nu(r_1)\frac{1}{|\,r_1 - r_2\,|}\phi_\delta^*(r_2)\phi_\lambda(r_2) \tag{2-6}$$

After the density matrix is converged through the iterative procedure of solving the Roothann-Hall Equation 2-3, the total energy is calculated by taking the sum over the electronic and nuclear energy.

$$E = \frac{1}{2}\sum_{\mu\nu}^{N} P_{\mu\nu}\left(H_{\mu\nu}^{core} + F_{\mu\nu}\right) + E_{nuc} \tag{2-7}$$

The SCF procedure starts with calculation of the core-Hamiltonian, the overlap matrix, the two-electron integrals and the initial guess of the density matrix, then the Fock matrix is constructed through Equation 2-4. The overlap matrix is first diagonalized and produces a transformation matrix. Using the transformation matrix, the Fock matrix is transformed through a similarity transformation. The MO coefficient matrix C and the orbital energies are obtained by diagonalizing the Fock matrix. The density matrix will be updated from C using Equation 2-2. This procedure is repeated until the density matrix is converged within a specified criterion and finally we can get the total electronic energy through Equation 2-7.

We have developed an in-house *ab initio* program named QUICK.[35] In this program the two-electron integral package is based on Obara and Saika's vertical recursion[36] and Head-Gordon and Pople's horizontal recursion[37] algorithms. Table 2-1 compares the timing of HF calculations using QUICK and GAMESS-US[38] on polyalanine systems $(ala)_n (n = 1,2,3,5,10,15)$. In general, the computational efficiency of QUICK is comparable to GAEMSS-US. We have implemented our linear scaling algorithms in the QUICK program.

Table 2-1. Average computational time for each SCF cycle of HF/6-311G** single point calculations (seconds).

System	Basis functions	QUICK(s)	GAMESS(s)	Ratio
$(\text{ala})_1$	137	6.9	4.4	1.6
$(\text{ala})_2$	262	38	26	1.4
$(\text{ala})_3$	387	97	72	1.3
$(\text{ala})_5$	637	298	223	1.3
$(\text{ala})_{10}$	1262	1366	1233	1.1
$(\text{ala})_{15}$	1887	3689	3981	0.9

2.2 Second-Order Møller-Plesset Perturbation Theory

With the second order Møller-Plesset perturbation theory, the electronic Hamiltonian is written as

$$H = H_0 + V \tag{2-8}$$

where H_0 is the Hartree-Fock Hamiltonian and the perturbation V is the difference between the electronic and Hartree-Fock Hamiltonian:

$$H_0 = \sum_i f(r_i) \tag{2-9}$$

$$V = \sum_{i<j} g(r_i - r_j) - \sum_i v^{HF}(r_i) \tag{2-10}$$

Expansion of the wave function and energy terms by Rayleigh-Schrödinger perturbation theory[1] gives the ground state electronic energy:

$$E = E^{(0)} + E^{(1)} + E^{(2)} + E^{(3)} + \cdots \tag{2-11}$$

Second order Møller-Plesset truncates this expansion at the $E^{(2)}$. The sum of the $E^{(0)}$ and $E^{(1)}$ is the Hartree-Fock energy. For simplicity, in this discussion of MP2 calculation we only consider the closed-shell (RHF) case and use spatial orbitals. The canonical MP2 electron correlation energy is expressed as[36,39]

$$E_{corr}^{(2)} = \sum_{ij}^{occ}\sum_{ab}^{vir}(ia \mid jb)\left[\frac{2(ia \mid jb) - (ib \mid ja)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}\right] \qquad (2\text{-}12)$$

where i, j are occupied molecular orbitals ψ_i, ψ_j with eigenvalues ε_i and ε_j and a,b are virtual molecular orbitals ψ_a, ψ_b with eigenvalues ε_a and ε_b. The $(ia \mid jb)$ term is the MO electron-repulsion integral (ERI) given by

$$(ia \mid jb) = \iint dr_1 dr_2 \psi_i^*(r_1)\psi_a(r_1) r_{12}^{-1} \psi_j^*(r_2)\psi_b(r_2) \qquad (2\text{-}13)$$

$$\psi_i = \sum_{\mu=1}^{N} C_{\mu i}\phi_\mu \qquad (2\text{-}14)$$

2.3 Ab Initio Linear-Scaling Methodology

2.3.1 Divide-and-Conquer Approach

2.3.1.1 DC-HF method

In protein systems, the divide-conquer approach is based on the chemical locality; this assumes that local regions of a protein are only weakly influenced by the atoms that are far away from the region of interest. The whole system is divided into fragments called core regions ($Core^\alpha$). A buffer region ($Buffer^\alpha$) is assigned for each core region to account for the environmental effects. The combination of every core region and its buffer region constitutes each individual subsystem (R^α) as illustrated in Figure 2-1. Local MOs of each subsystem are solved by Roothaan-Hall equation

$$F^\alpha C^\alpha = S^\alpha C^\alpha E^\alpha \qquad (2\text{-}15)$$

where F^α and S^α are local Fock matrix and local overlap matrix, respectively.

$$F_{\mu\nu}^\alpha = \begin{cases} F_{\mu\nu} & \text{if } \chi_\mu \in R^\alpha \text{ and } \chi_\nu \in R^\alpha \\ 0 & \text{elsewhere} \end{cases} \qquad (2\text{-}16)$$

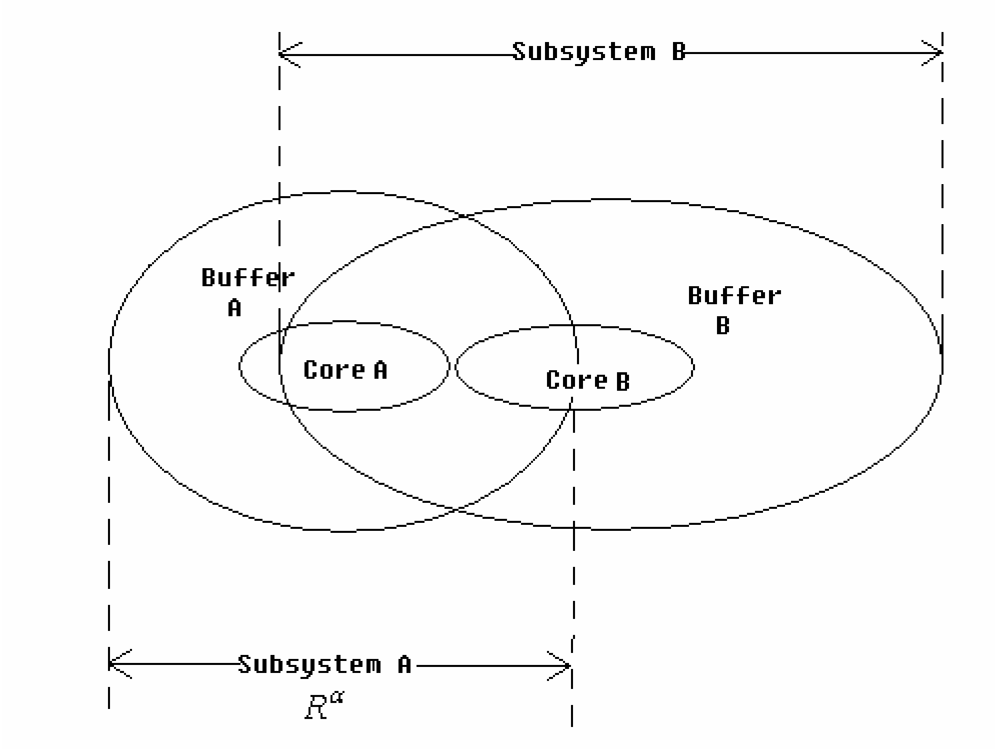


Figure 2-1. Graphical representation of the subsetting scheme used in divide-and-conquer calculations.

After the local MO coefficient matrices C^{α} are obtained, the total density matrix of the whole system is given by

$$P_{\mu\nu} = \sum_{\alpha=1}^{N_{sub}} P_{\mu\nu}^{\alpha} = \sum_{\alpha=1}^{N_{sub}} D_{\mu\nu}^{\alpha} p_{\mu\nu}^{\alpha} \tag{2-17}$$

where $D_{\mu\nu}^{\alpha}$ is the partition matrix, (see Figure 2-2)

$$D_{\mu\nu}^{\alpha} = \begin{cases} 1 & \phi_{\mu} \in Core^{\alpha} \quad \text{and}\, \phi_{\nu} \in Core^{\alpha} \\ 1/2 & \phi_{\mu} \in Core^{\alpha} \quad \text{and}\, \phi_{\nu} \in Buffer^{\alpha} \;\text{or}\; \phi_{\mu} \in Buffer^{\alpha} \quad \text{and}\, \phi_{\nu} \in Core^{\alpha} \\ 0 & \phi_{\mu} \notin Core^{\alpha} \quad \text{and}\, \phi_{\nu} \notin Core^{\alpha} \end{cases} \tag{2-18}$$

and $p_{\mu\nu}^{\alpha}$ is the local density matrix defined by

$$p_{\mu\nu}^{\alpha} = \sum_{i}^{LMOs} n_i^{\alpha} C_{\mu i}^{\alpha} C_{\nu i}^{\alpha *} \tag{2-19}$$

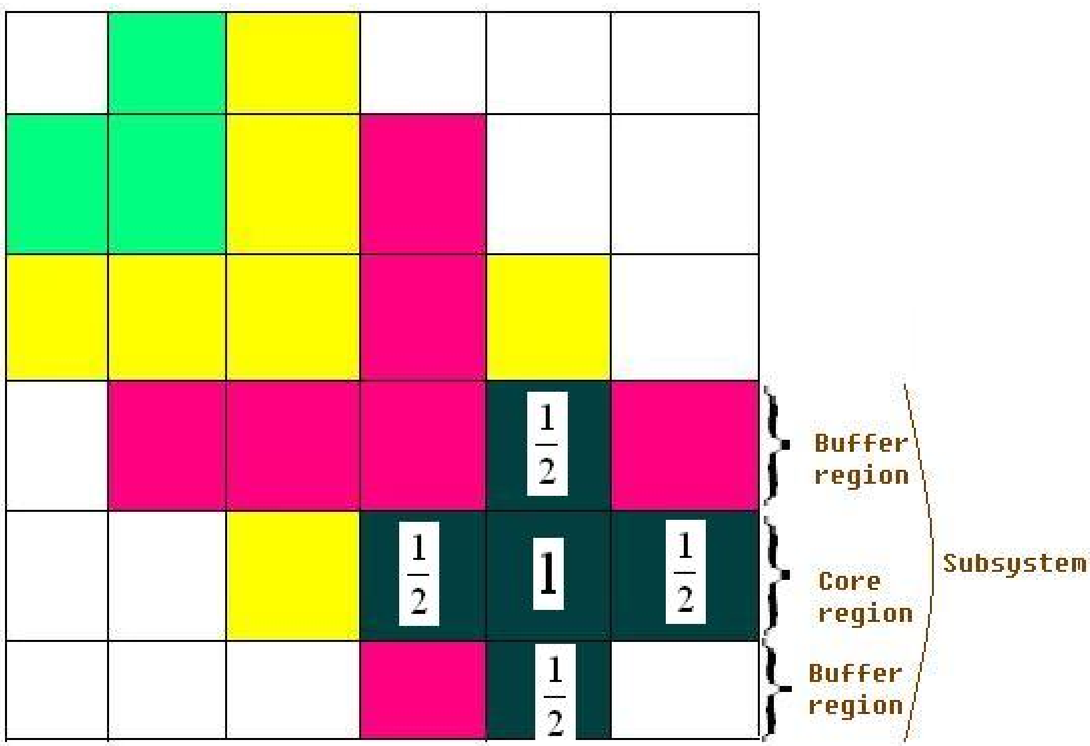


Figure 2-2. The way to assemble the full density matrix from the density matrices of individual subsystems based on the divide-and-conquer approach.

where n_i^α is a smooth approximation to the Heaviside step function:

$$n_i^\alpha = \frac{2}{1+\exp[(\varepsilon_i^\alpha - \varepsilon_F)/kT]} \tag{2-20}$$

ε_F is determined through the normalization of the total number of electrons of the whole system.

$$N_{elec} = \sum_\alpha \sum_\mu (P^\alpha S^\alpha)_{\mu\mu} \tag{2-21}$$

After the density matrix is converged, the total HF energy is given as

$$E_{HF}^{DC} = \frac{1}{2}\sum_\alpha \sum_{\mu\nu} P_{\mu\nu}^\alpha (H_{\mu\nu}^\alpha + F_{\mu\nu}^\alpha) \tag{2-22}$$

where $H_{\mu\nu}^\alpha$ is the local one-electron core Hamiltonian matrix similar to the definition of local Fock matrix in Equation 2-16.

For HF calculations on large systems, the construction of the Coulomb matrix and exchange matrix along with the diagonalization of the Fock matrix constitute the three most time-consuming steps. The Hamiltonian matrix diagonalization intrinsically scales as O(N^3). In the divide-and-conquer scheme the diagonalization calculation is performed on each submatrix, which will naturally make the SCF diagonalization step scale linearly with the number of subsystems. However, it is important to realize that the divide-and-conquer algorithm does not help to reduce the scale of computation of the Coulomb matrix and exchange matrix. The continuous fast multipole method (CFMM)[8,10-12,40-43] and the linear exchange K approach (LinK)[44,45] provide ways in which the Coulomb matrix and exchange matrix can be made linear-scaling, respectively.

2.3.1.2 DC-MP2 method

If we only perform the partial transformation from AO $\phi_\mu, (\mu = 1,2,\cdots N)$ to the first MO i , the MO ERI will be

$$(ia \mid jb) = \sum_{\mu} C_{\mu i} (\mu a \mid jb) \tag{2-23}$$

In the divide-conquer approach, the buffer regions are overlapped, thus, we can not simply sum the electron correlation energy of each subsystem. To eliminate the double counting of the correlation energy contributed from buffer regions, we employ the correlation energy decomposition scheme proposed by Nakai and co-workers[20,21]. MO ERI is decomposed to each core region ($Core(\alpha)$) when we transfer the AO $\phi_\mu, (\mu = 1,2,\cdots N)$ to MO i

$$(ia \mid jb) = \sum_{\alpha} \sum_{\mu \in Core(\alpha)} C_{\mu i} (\mu a \mid jb) \tag{2-24}$$

Then the correlation energy is given by

$$E_{corr}^{(2)} = \sum_{\alpha} \sum_{ij}^{occ} \sum_{ab}^{vir} \sum_{\mu \in Core(\alpha)} C_{\mu i} (\mu a \mid jb) \left[\frac{2(ia \mid jb) - (ib \mid ja)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \right] \tag{2-25}$$

The total correlation energy can be approximated by taking the sum of the correlation of each subsystem.[20]

$$E_{corr}^{(2)} = \sum_{\alpha}^{subsystem} E_{corr}^{\alpha} \tag{2-26}$$

$$E_{corr}^{(2)} = \sum_{\alpha}^{subsystems} \sum_{i^{\alpha} j^{\alpha}}^{occ(\alpha)} \sum_{a^{\alpha} b^{\alpha}}^{vir(\alpha)} \sum_{\mu \in Core(\alpha)} C_{\mu i}^{\alpha} (\mu a^{\alpha} \mid j^{\alpha} b^{\alpha}) \left[\frac{2(i^{\alpha} a^{\alpha} \mid j^{\alpha} b^{\alpha}) - (i^{\alpha} b^{\alpha} \mid j^{\alpha} a^{\alpha})}{\varepsilon_i^{\alpha} + \varepsilon_j^{\alpha} - \varepsilon_a^{\alpha} - \varepsilon_b^{\alpha}} \right] \tag{2-27}$$

where i^{α}, j^{α} are the local occupied molecular orbitals $\psi_i^{\alpha}, \psi_j^{\alpha}$ of the subsystem α which have their eigenvalues ε_i^{α} and ε_j^{α} lower than the fermi energy ε_F, and a^{α}, b^{α} are virtual molecular orbitals $\psi_a^{\alpha}, \psi_b^{\alpha}$ of the subsystem α with eigenvalues ε_a^{α} and ε_b^{α} higher than ε_F.

In DC-MP2 calculations, the evaluation of the subsystem correlation energy using Equation 2-27 scales as $O(m^5)$, where *m* denotes the number of basis functions of each subsystem. Nevertheless, the size of the subsystem is independent of the size of the entire system. The total computational cost would be $\sim O(Nm^5)$, where *N* is the number of the subsystems. Therefore, the calculation on correlation energy scales linearly for large molecules. Another advantage of the DC-MP2 approach is that the memory usage is also independent of the size of the whole system. The maximum memory requirement is only decided by the largest subsystem.

2.3.2 FMO Approach

The fragment-based approach FMO has already been applied to MP2 theory and is capable to deal with macromolecules within a reasonable computational cost. The

FMO computational procedure is as follows[46,47]: first, the protein is divided into N fragments containing one or two amino acid residues each. The electronic structure of a single fragment (monomer) is solved in the external coulomb field contributed by the remaining $(N-1)$ monomers repeatedly until all the density matrices of the monomers are self-consistent. Secondly, the energy of every fragment pair (dimer) is solved in an approximate electrostatic field generated by the remaining $(N-2)$ monomers. The energy of each trimer can be calculated in the same way. Finally, the total energy of the protein is obtained using the following expression (higher order many-body interaction energies are neglected):

$$\begin{aligned} E_{FMO}^{Total} &= \sum_{I=1}^{N} E_I + \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} E_{IJ} - E_I - E_J + \sum_{I=1}^{N-2} \sum_{J=I+1}^{N-1} \sum_{K=J+1}^{N} \{(E_{IJK} - E_I - E_J - E_K) \\ &-(E_{IJ} - E_I - E_J) - (E_{JK} - E_J - E_K) - (E_{KI} - E_K - E_I)\} \end{aligned} \tag{2-28}$$

where N represents the number of fragments. In our implementation, we take two consecutive amino acid residues as a fragment. E_I , E_{IJ} and E_{IJK} are the monomer, dimer and trimer energies, respectively. Because of the computational cost, we truncated the energy contributions after the two-body expansion (termed as FMO2). As shown in a previous study,[48] the deviation between FMO2-MP2 computed correlation energies and full MP2 calculations, on several model protein systems, is ~2.1kcal/mol. Thus, FMO2 is a practical approach that strikes a compromise between accuracy and computational expense in studies of macromolecules. In FMO2 expansion, the restricted Hartree-Fock (RHF) energy and the MP2 correlation energy are obtained similarly to Equation 2-28

$$E_{FMO2}^{RHF} = \sum_{I=1}^{N} E_I^{RHF} + \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} (E_{IJ}^{RHF} - E_I^{RHF} - E_J^{RHF}) \tag{2-29}$$

$$E_{FMO2}^{corr} = \sum_{I=1}^{N} E_I^{corr} + \sum_{I=1}^{N-1} \sum_{J=I+1}^{N} (E_{IJ}^{corr} - E_I^{corr} - E_J^{corr}) \quad (2\text{-}30)$$

where E_I^{corr}, E_{IJ}^{corr} are the MP2 correlation energy of the monomer and dimer, respectively. By adding the MP2 electron correlation energy to the FMO2-HF energy, we obtain the FMO2-MP2 energy:

$$E_{FMO2}^{MP2} = E_{FMO2}^{RHF} + E_{FMO2}^{corr} \quad (2\text{-}31)$$

2.3.3 MFCC Approach

The basic idea of the MFCC approach[49] is similar to the original divide-and-conquer approach of Yang[16] but differs in technical treatment. The MFCC method also has some features in common with the fragment molecular orbital method (FMO) of Kitaura *et al.*[50] in that the protein is partitioned into amino-acid fragments. However, detailed treatment of protein fragment is significantly different in both approaches. The MFCC approach has been successfully applied to a range of problems including protein–water,[51] protein–ligand systems,[52,53] and protein–ligand geometry optimization.[54]

Using the MFCC approach[55] (illustrated in Figure 2-3), the total electron density of a long polymer such as protein with N-amino acids can be obtained by linear combination of individual densities of various capped fragments by an MFCC ansatz[56]

$$\rho = \sum_{k=1}^{N} \rho_k - \sum_{k=1}^{N-1} \rho_k^{cc} - \sum_{k=1}^{N_d} \rho_k^{dc} \quad (2\text{-}32)$$

where ρ_k is the density of the *k*th protein fragment, ρ_k^{cc} is the density of the *k*th concap (conjugate caps), ρ_k^{dc} is the density of the *k*th disulfide concap (if any), and N_d is the number of disulfide bonds in protein which are cut in the MFCC approach.[57] The same result can be obtained for the electrostatic potential and dipole moment.[56]

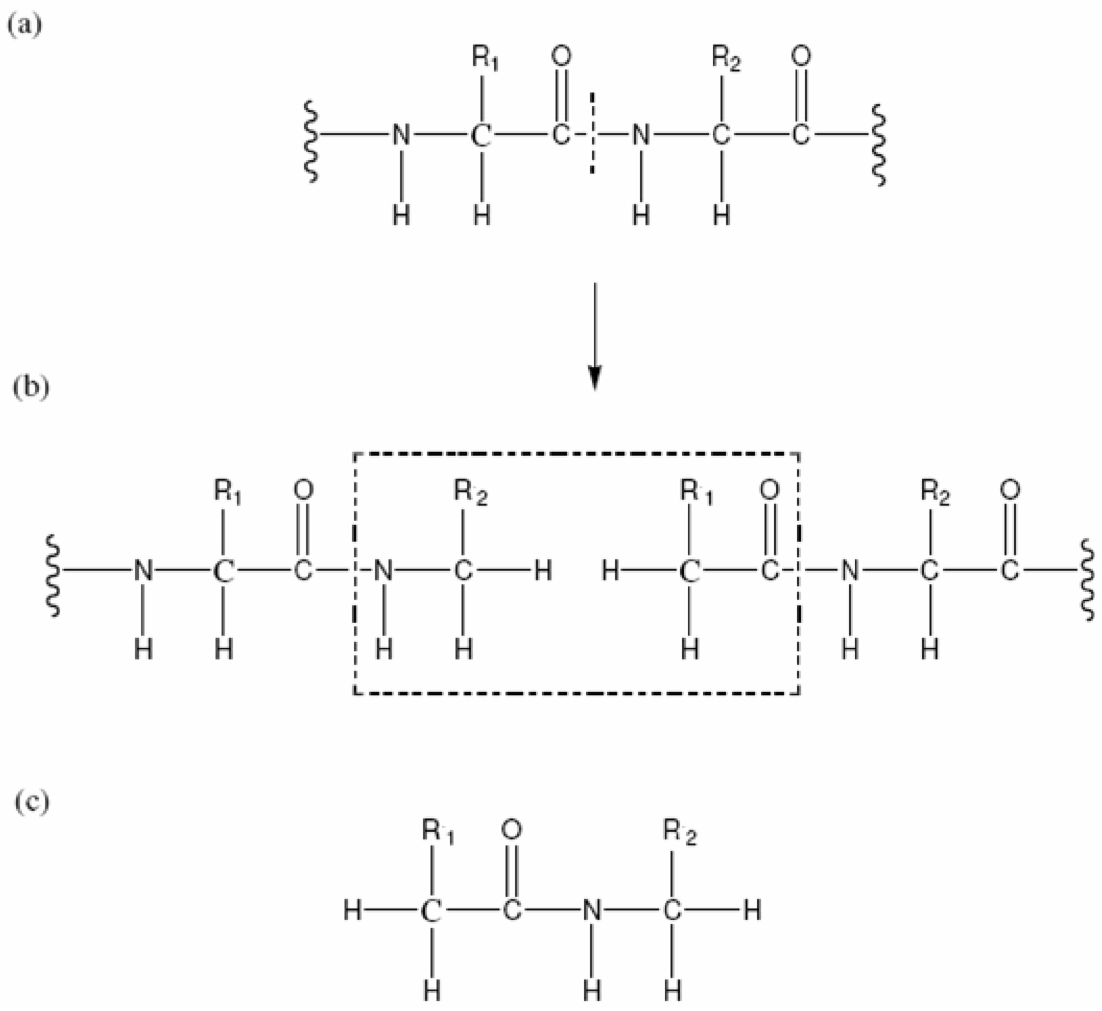


Figure 2-3. The MFCC scheme in which the peptide bond is cut (a) and the fragments are capped with C_{cap} and its conjugate C_{cap}^{*} (b). The atomic structure of the concap is shown in (c). The concap is defined as the fused molecular species of C_{cap}^{*} –C_{cap}.

It is straightforward to verify that the total electron density obtained from Equation 2-32 is correctly normalized

$$\int \rho dr = \sum_{k=1}^{N} \int \rho_k dr - \sum_{k=1}^{N-1} \int \rho_k^{cc} dr - \sum_{k=1}^{N_d} \int \rho_k^{dc} dr = \sum_{k=1}^{N} N_k - \sum_{k=1}^{N-1} N_k^{cc} - \sum_{k=1}^{N_d} N_k^{dc} = N_{total} \qquad \text{(2-33)}$$

where N_k and N_k^{cc} are, respectively, the number of electron of the *k*th protein fragment (capped) and concap, and N_k^{dc} is the number of electron of the *k*th disulfide concap. Thus the density in Equation 2-32 automatically has the correct normalization.

After the density of the full protein system is obtained from the MFCC calculation, we can employ density functional theory (DFT) to compute total energy of protein *E* by the DFT energy expression

$$E[\rho] = T[\rho] - \sum_{\alpha} Z_{\alpha} \int \frac{\rho(r)}{|R_{\alpha} - r|} dr + \frac{1}{2} \int \phi(r) \rho(r) dr + \sum_{\alpha,\beta} \frac{Z_{\alpha} Z_{\beta}}{R_{\alpha\beta}} + E_{xc}[\rho] \tag{2-34}$$

where $T[\rho]$ is the kinetic energy, $\phi(r)$ is the electrostatic potential (electron contribution only), and $E_{xc}[\rho]$ is the exchange-correlation energy.

Since the analytical form of the kinetic energy functional $T[\rho]$ is unknown, we make a new MFCC ansatz for a two component A–B system treated with the MFCC approach

$$T[\rho_{AB}] = T[\rho_A] + T[\rho_B] - T[\rho^{cc}] \tag{2-35}$$

where $T[\rho_{AB}]$ is the kinetic energy of the A–B system, $T[\rho_A]$ and $T[\rho_B]$ are, respectively, the kinetic energy of the capped A and B fragments, and $T[\rho^{cc}]$ is the electron density of the concap species. It is easy to verify that the above relation would be exact if any of the caps (C or C*) includes the complete counter part of the system. Equation 2-35 is easily generalized to an *N*-component system like protein (assuming no disulfide bond).

$$T[\rho] = \sum_{k=1}^{N} T_k[\rho_k] - \sum_{k=1}^{N-1} T_k^{cc}[\rho_k^{cc}] \tag{2-36}$$

Thus the total kinetic energy of the system can be obtained by simple combination of kinetic energies of individual fragments, albeit approximately.

The calculation of the other energy terms in Equation 2-34 is done as follows. The potential energy (PE) (second term) In Equation 2-34 can be obtained in a straightforward fashion

$$PE = -\sum_{\alpha} \int \frac{Z_{\alpha}}{|R_{\alpha} - r|} \rho(r) dr = \sum_{\alpha} Z_{\alpha} \phi(R_{\alpha}) \tag{2-37}$$

where $\phi(R_{\alpha})$ is the total electrostatic potential (electronic contribution only) obtained from MFCC calculation similar to Equation 2-32 evaluated at the nuclear center of atom α .

The evaluation of the Coulomb energy (EE) (third term in Equation 2-34) is done by numerical integration. In order to reduce errors in numerical integration, we use the following strategy to perform numerical integration

$$\begin{aligned} EE = \frac{1}{2} \int [\phi(r)\rho(r) - [\sum_{k=1}^{N} \phi_k(r)\rho_k(r) - \sum_{k=1}^{N-1} \phi_k^{cc}(r)\rho_k^{cc}(r)]]dr + \\ \frac{1}{2} \int [\sum_{k=1}^{N} \phi_k(r)\rho_k(r) - \sum_{k=1}^{N-1} \phi_k^{cc}(r)\rho_k^{cc}(r)]dr = EE(1) + EE(2) \end{aligned} \tag{2-38}$$

The second term in the above equation can be obtained by simple combination from MFCC calculation

$$EE(2) = \sum_{k=1}^{N} EE_k - \sum_{k=1}^{N-1} EE_k^{cc} \tag{2-39}$$

where EE_k and EE_{k}^{cc} are the Coulomb energy of individual fragments and concap that can be obtained directly from Gaussian calculation for each individual fragment. Becke's method[58] is employed in numerical integration for EE(1). Using the above scheme, the numerical integration error of EE is significantly reduced. Since the exchange and correlation energies are relatively small compared to other energy terms in Equation 2-34, Becke's integration method is applied directly to evaluate exchange (Ex) and correlation (Ec) energies.

2.3.4 AF-QM/MM Approach

Figure 2-4 shows the subsetting scheme in the current AF-QM/MM implementation. The entire system is divided into non-overlapping fragments termed core regions. The residues in a certain range from the core region are assigned as the buffer region. Both the core region and its buffer region are treated by quantum mechanics, while the rest of the system is described by an empirical point charge model. The purpose of the buffer area is to include the local QM effects on the chemical shifts. Each fragment-centric QM/MM calculation is carried out separately. Only the shielding constants of the atoms in the core region are extracted from the individual QM/MM calculation. A more detailed illustration of the automated fragmentation scheme is presented in Figure 2-5. In this chapter, each residue is taken as the core region. To preserve the electron delocalization across the peptide bond, we adopt a different definition of the residue which consists of the –CO-NH-CHR- atoms as shown in Figure 2-5a. We introduce a generalized molecular cap to take into account the QM polarization effect and charge transfer within the first shell from the residue of interest as shown in Figure 2-5b. The concept of the generalized molecular cap is an extension of the molecular conjugate fractionation with conjugate caps approach (MFCC).[55,59] Only the sequentially connected residues are included in the molecular caps for the standard MFCC approach. Here we extend the molecular cap to non-bonded residues which have hydrogen bonding interactions, ring current effects and other QM effects in the vicinity of the core region. The non-neighboring residues in the buffer region are simply capped by hydrogen atoms to construct the closed-shell fragment. The position of the additional hydrogen atom is determined in the same fashion as the MFCC method.[55] In this dissertation, we adopt the following distance-dependent criterion to include residues into the buffer region of each core residue. (1) If one atom of the residue outside the core region is less than 4Å away from any atom in the core region, and at least one of the two atoms is a non-hydrogen atom. (2) If the distance between one hydrogen atom in the core region and the other hydrogen atom outside the core region is less than 3Å away from each other. (3) If a heavy atom on an aromatic ring is within 5Å from any atom in the core region. Of course, other distance-dependent criterion could be used to further optimize the choice of the buffer region.

The remaining atoms beyond the buffer region are treated by molecular mechanics. A point charge model is employed to account for the empirical electrostatic field outside the QM region. We use the full point charges for those junction atoms which are replaced by hydrogen atoms. Since a buffer region is added to smoothly link the core region and MM environment, atoms on the boundary between the QM and MM regions are relatively far apart from the core region and their influence is attenuated. Other approaches such as the field-adapted adjustable density matrix assembler (FA-ADMA)[60] method and the generalized energy-based fragmentation approach (GEBF) [61] use similar treatments for the interaction between distant residues. In this chapter, we are not aiming to obtain the total energy of the protein. Our purpose is to develop a more generalized automated fragmentation approach to accurately calculate NMR chemical shifts. By using a general criterion to assign a buffer zone to each residue, we can reduce the size of each fragment in order to make the QM calculation as small as possible until we strike a compromise between the desired accuracy and the computational cost.

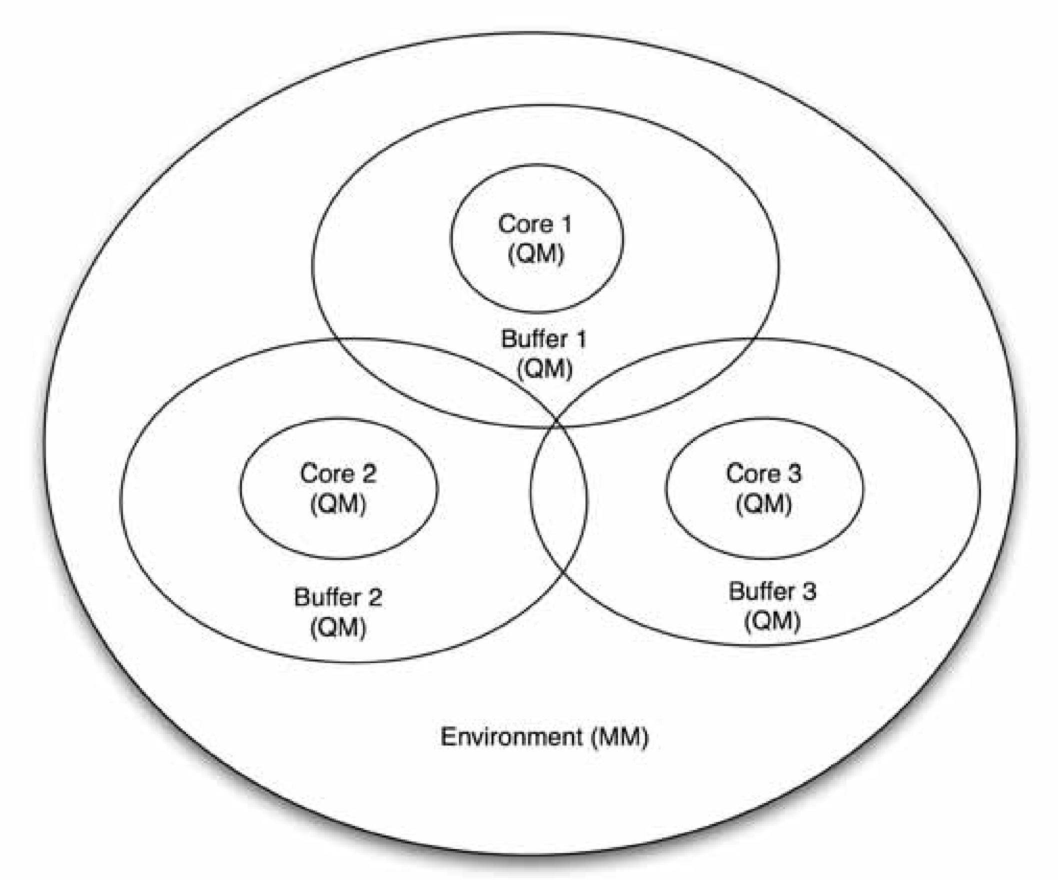


Figure 2-4. The subsetting scheme for the automated fragmentation AF-QM/MM approach.

Although the total number of residue pairs is proportional to the square of the number of residues, the size of each fragment is independent of the overall protein size because each residue can only have limited number of residues in its vicinity. Hence, the largest fragment normally contains less than 250 atoms consisting of C, H, O, N, and S, which is an affordable calculation at the HF and DFT level.

The idea of using partial MM charges is borrowed from the popular QM/MM approach except that the current AF-QM/MM scheme is applied to the entire protein system. The AF-QM/MM method has a number of attractive features. First, the

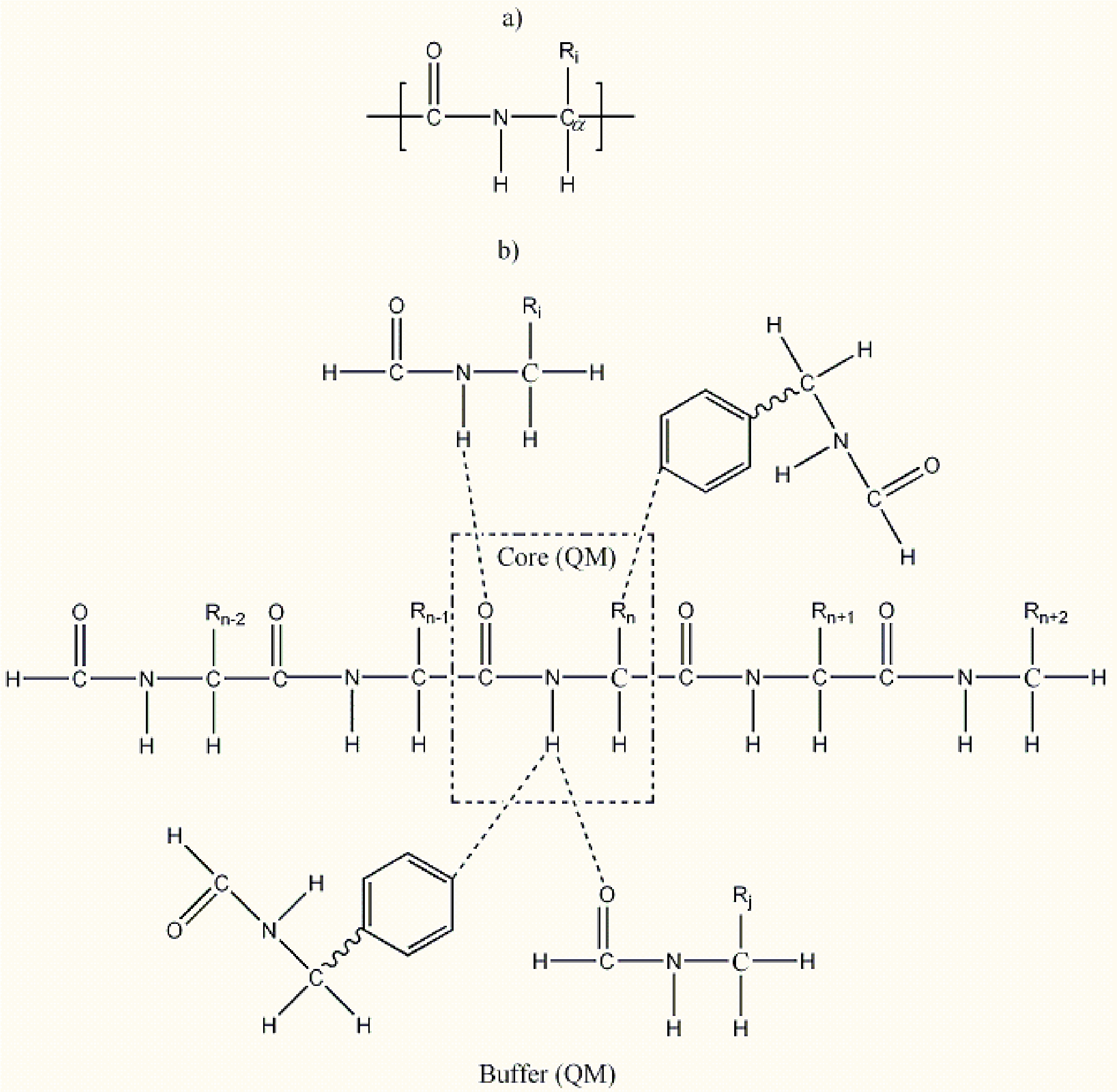


Figure 2-5. a) The definition of the residue unit used in this chapter. b) The n*th* amino acid is the core region. Sequentially connected (n-2)*th*, (n-1)*th*, (n+1)*th* and (n+2)*th* residues are included in the buffer region. In addition, the residues in spatial contact with the n*th* residue are also assigned to the buffer region (see text for further details).

construction of the density matrix or Hamiltonian of the entire molecular system is avoided. All the fragment-centric QM/MM calculations are mutually independent and parallelizable. Secondly, there is no need to diagonalize the full Hamiltonian matrix which is the bottleneck in linear-scaling calculations of macromolecules. Thirdly, the memory requirement only depends on the largest size of the divided fragment and does not increase with the size of the entire system. Fourthly, this approach can be extended beyond HF and DFT, to high-level electron-correlated *ab initio* methods such as second-order Møller-Plesset perturbation theory (MP2) or Coupled-Cluster theory (CC) if so desired.[62-64]

CHAPTER 3
DIVIDE-AND-CONQUER HARTREE-FOCK AND MP2 CALCULATIONS ON PROTEINS

3.1 Introduction

The state-of-the-art linear-scaling algorithms, which make the computational cost scale linearly $O(N)$ with the system size, have attracted the focus of many theorists during the past decade.[6-9,12,65-67] Much effort has been devoted to the development of linear-scaling methods in order to compute the total energy of large molecular systems at the Hartree-Fock (HF) or density functional theory (DFT) level.[6,9,12,16,17,42,68,69] One of the challenges is to speed up the calculation of long-range Coulomb interactions when assembling the Fock matrix elements. Fast multipole based approaches have successfully reduced the scaling in system size to linear[8,12,42,66,67] and made HF and DFT calculations affordable for larger systems when small to moderate sized basis sets are utilized. The more recently developed Fourier Transform Coulomb method of Fusti and Pulay[70,71] reduced the steep $O(N^4)$ scaling in basis set size to quadratic and makes the calculations much more affordable with larger basis sets.[72] There is also a class of fragment-based methods for quantum calculation of protein systems including the divide and conquer (D&C) method of Yang[16], Yang and Lee,[17] Dixon and Merz,[18] Gogonea *et al.*,[73] Shaw and St-Amant,[22] and Nakai and co-workers,[74-77] the adjustable density matrix assembler (ADMA) approach method of Exner and Mezey,[60,69,78,79] the fragment molecular orbital (FMO) method of Kitaura and co-workers,[7,46,47] and the molecular fractionation with conjugate caps (MFCC) approach developed by Zhang and co-workers.[49,55] Most applications of these methods to protein systems have been largely limited to semiempirical, HF and DFT calculations. Among these approaches, FMO has been applied to higher level *ab initio* calculations such as second-order Møller-Plesset perturbation theory (MP2)[48] and coupled cluster theory (CC).[80] Nakai and co-workers have recently proposed DC-MP2[20,74,77] and DC-CCSD[21] approaches; however, only systems of linear chains or near-linear chains have been tested so far for the divide-and-conquer algorithm for *ab initio* calculations.

The aim of our current research is to further develop and validate the divide-and-conquer (DC)[16-22] methodology to aid in the application of *ab initio* methods to biomacromolecules. In this study, our goal is to validate divide-and-conquer algorithm for Hartree-Fock calculations on globular proteins. Moreover, we propose a fragment-based initial guess using molecular fractionation with conjugated caps (MFCC) method

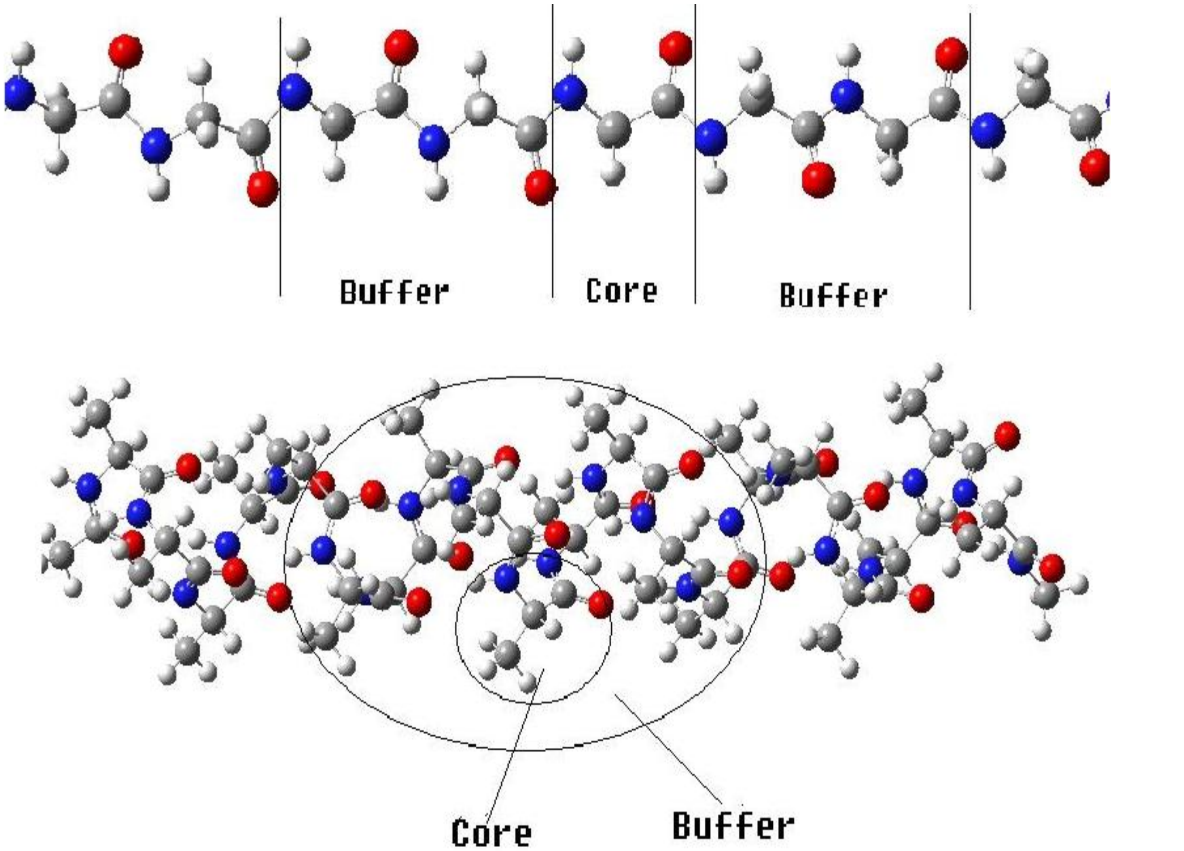


Figure 3-1. The subsetting schemes for divide-and-conquer calculations on the extended polyglycine $(\mathrm{Gly})_n$ (upper) and polyalanine in an $\alpha-$helical structure ($\alpha-(\mathrm{Ala})_n$, bottom).

to reduce the number of SCF cycles, and different division schemes are compared.

3.2 Accuracy and Timing Comparisons

3.2.1 DC-HF Calculations

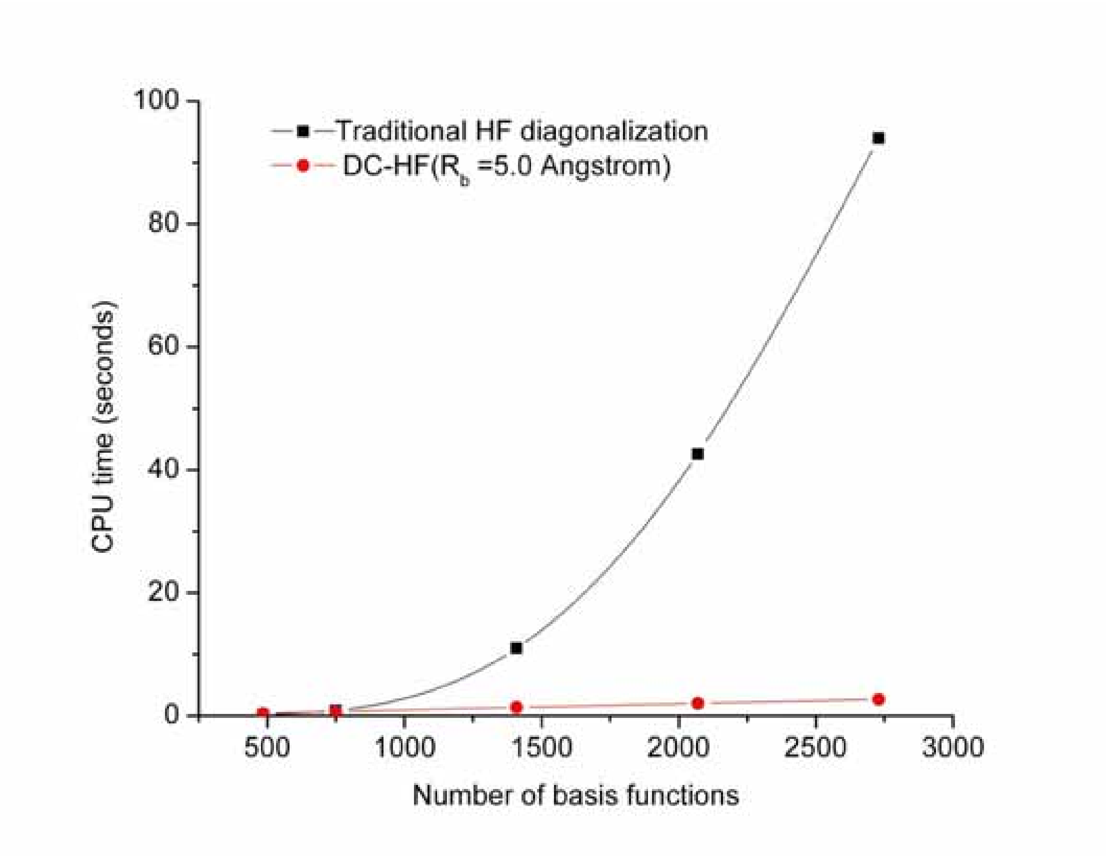


Figure 3-2. The average computational time to diagonalize the Fock matrix in each SCF cycle using traditional HF and DC-HF for a series of extended polyglycines at the HF/6-31G* level.

In this section, we assess the DC-HF approach performance by performing calculations on two types of simple systems: extended polyglycine $(\text{gly})_n$ and an alpha-helix of polyalanine ($\alpha-(\text{ala})_n$ see Figure 3-1). All the calculations discussed here use the 6-31G* basis set. A buffer radius of $R_b = 5.0$ Å was adopted for all DC-HF calculations. Within this definition we include all the residues that contain any atoms within 5Å from the core region as part of the buffer region. A comparison of the CPU

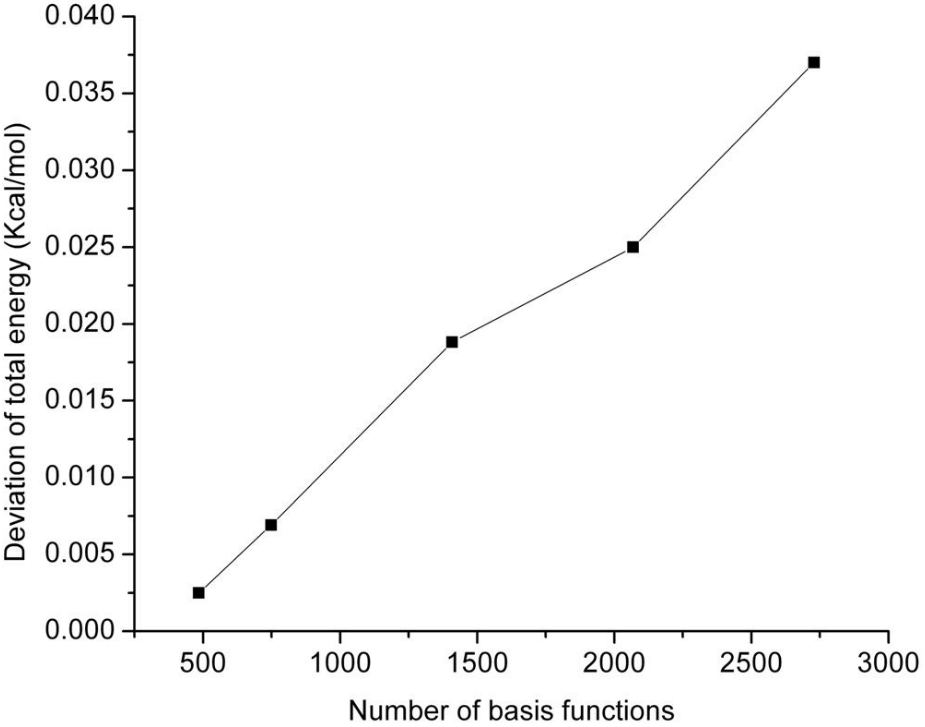


Figure 3-3. The accuracy of the total energy calculated by the DC-HF approach on extended polyglycine systems compared to full system calculations.

time required to solve the SCF equations on the extended polyglycine $(gly)_n$ (n=6~40) using the standard HF and DC-HF approaches is shown in Figure 3-2. As expected, the computational scale for the DC-HF diagonalization calculation is O(N), in contrast to $O(N^{2.9})$ for the traditional HF SCF diagonalization on the full Fock matrix of the entire system. Moreover, since the polyglycine is extended, the basis set crossover point is between 485 and 749. Figure 3-3 shows the deviation of DC-HF energy compared to the full system calculation on extended polyglycine systems. The error becomes larger as the size of the system increases; however, all of the deviations are within 0.04 kcal mol^{-1}. This good accuracy suggests that we can employ the divide-and-conquer scheme to study large, 3-dimensional systems. The computational cost and accuracy of DC-HF

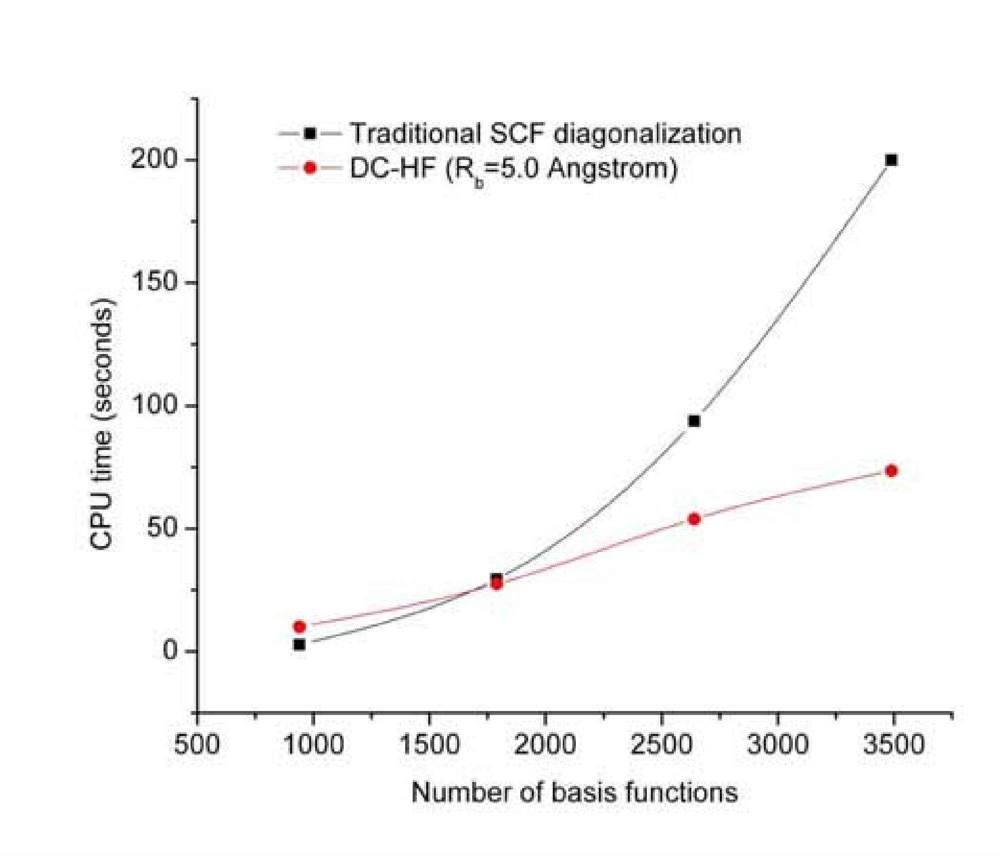


Figure 3-4. Similar to Figure 3-2, but for the polyalanine systems in an $\alpha -$helical structure $\alpha -$(Ala)$_n$.

for $\alpha -$(ala)$_n$ (n=10~40) systems are illustrated in Figures 3-4 and 3-5, respectively. Because of the helix structure, each subsystem contains a larger number of residues than in the extended system using the same buffer size. As illustrated in Figure 3-4, the crossover point is around 1789, which is over 2-times larger than for the polyglycine example. Each DC-HF diagonalization SCF cycle in this example scales as $O(N^{1.1})$, in contrast to $O(N^{2.7})$ for the traditional HF diagonalization cost. Furthermore, the total energy errors for the α -helical polyalanines are slightly larger than those for the extended polyglycine systems (see Figure 3-5), but they are still in a good agreement

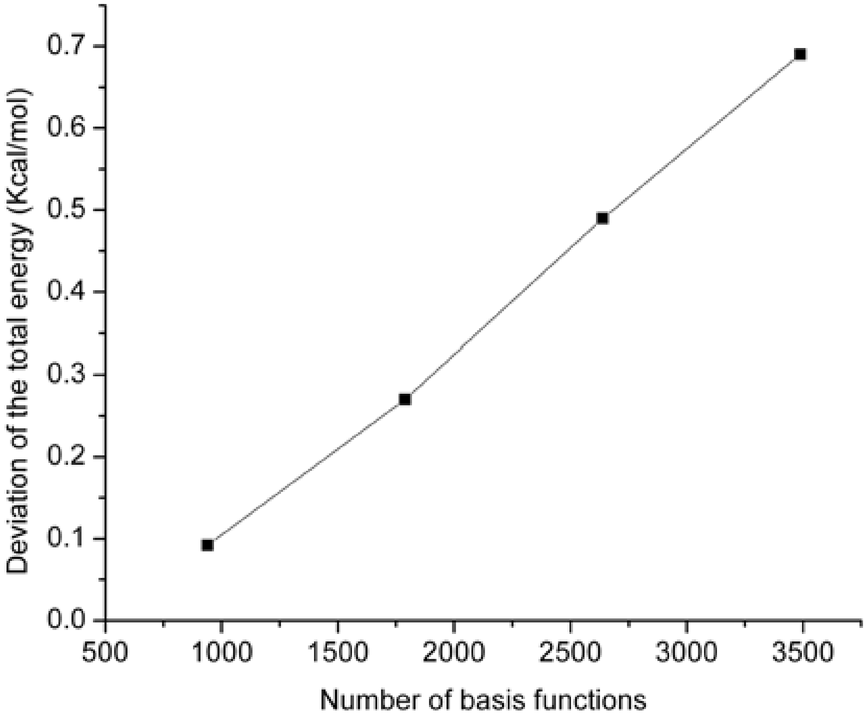


Figure 3-5. Similar to Figure 3-3, but for the polyalanine systems in an $\alpha -$helix structure $\alpha -(\text{Ala})_n$.

with the full system calculations since the largest error is less than 0.7 kcal mol^{-1} for $\alpha -(\text{ala})_{40}$.

In the current DC-HF approach, the scale for the computation of the Coulomb matrix is still $O(N^2)$ after prescreening the two-electron integrals. When we apply Equation 2-16 to construct the subsystem Fock matrix, the long-range Coulomb interactions between the local subsystem and distant atoms cannot be circumvented, thus, it should be emphasized that the D&C algorithm itself does not reduce the scale of Coulomb and exchange matrix evaluations and other approaches are necessary to achieve this result (*e.g.*, CFMM)[8,65-67].

3.2.2 DC-MP2 Calculations

We have also chosen extended polyglycines (gly)$_n$ as our test systems to validate the DC-MP2 approach. Figure 3-6 shows the computational cost as a function of the number of basis functions using two different buffer radii, specifically 3Å and 5Å. In contrast to the canonical MP2 calculations which scale as $O(N^{3.58})$, the DC-MP2 scales as $O(N^{1.34})$ with a buffer radius 3Å and $O(N^{1.25})$ with a buffer radius 5Å. DC-MP2 scales near linearly, but the CPU time of DC-MP2 with buffer size 5 Å has a larger prefactor than that with buffer size 3Å, because each subsystem is larger for a bigger buffer radius. However, the correlation energy calculated with a buffer size 5Å achieves

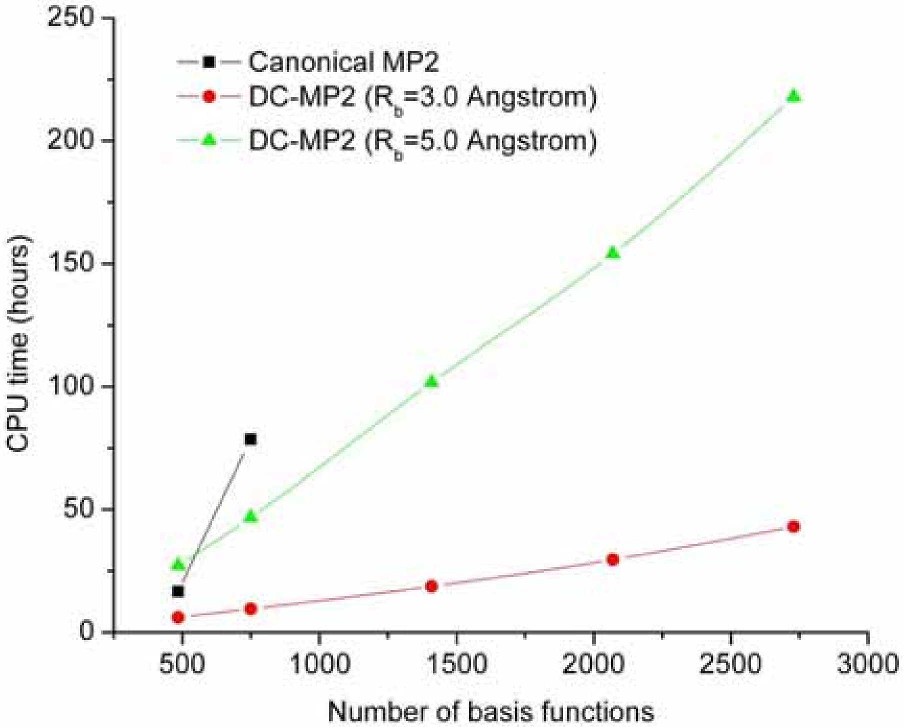


Figure 3-6. The comparison of the computational times between canonical MP2 and DC-MP2 for series of extended polyglycines $(gly)_n$ using 6-31G* basis set. Here, two different buffer sizes 5Å and 3Å are employed, respectively.

much more accurate results than for buffer size 3 Å (see Figure 3-7). For 1409 basis functions, the error of the total correlation energy for DC-MP2 calculation is only 0.08 kcal/mol with buffer size 5Å. As the buffer size increases it would be expected that a more accurate correlation energy would be obtained; the buffer size of 5Å strikes the compromise between the observed computational expense and attained accuracy.

The correlation energy decomposition scheme can be further applied to higher-order MP method and Couple Cluster theory (CC).[21] Our future development will also focus on divide-and-conquer scheme with higher level electron correlation methods.

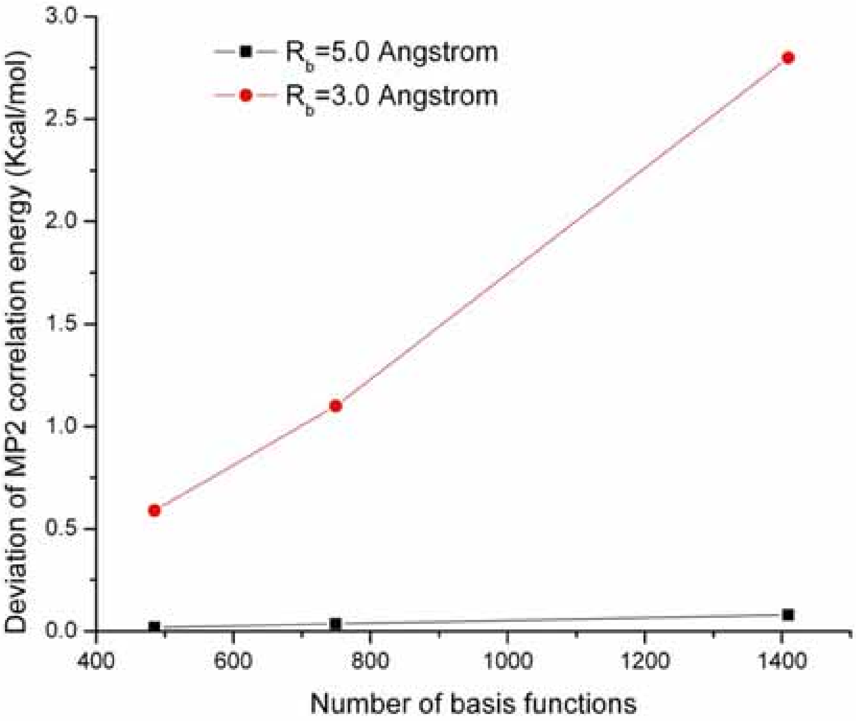


Figure 3-7. The errors of the DC-MP2 electron correlation energies as a function of the number of basis functions for the extended polyglycines $(gly)_n$ compared to full system calculaitons. Here, two different buffer sizes 5Å and 3Å are employed, respectively.

3.3 MFCC Initial Guess for Div&Con HF Calculations

Here we introduce a fragment-based initial guess for *ab initio* calculations using the molecular fractionation with conjugate caps (MFCC) algorithm as described elsewhere[55,81,82]. In the spirit of the MFCC approach, the full density matrix of the protein can be assembled by a linear combination of fragment density matrices

$$P_{\mu\nu} = \sum_{i=1}^{N_f} P_{\mu\nu}^{f}(i) - \sum_{j=1}^{N_c} P_{\mu\nu}^{cc}(j) \qquad (3\text{-}1)$$

where $P_{\mu\nu}^{f}(i)$ is the density matrix element of the *i*th protein fragment, $P_{\mu\nu}^{cc}(j)$ is the density matrix element of the *j*th conjugate cap. N_f and N_c are the total number of the protein fragments and conjugate caps, respectively. The MFCC partition scheme to cut a protein into amino-acid fragments and conjugate caps is shown in Figure 2-3. First, a series of single point HF calculations on all the fragments and conjugate caps are performed, then the full density matrix of the protein obtained using the converged fragment density matrices based on Equation 3-1 is taken as the initial guess for the subsequent divide-and-conquer HF calculations. All the *ab initio* calculations were implemented in our in-house developed quantum chemistry package QUICK.[35]

We have compared the number of SCF cycles necessary to reach convergence when the SAD (superposition of atomic densities) and MFCC initial guesses are used in the divide and conquer scheme using the 6-31G* basis set (see Table 3-1). The convergence criterion in all examples was set to 10^{-6} a.u. on the root-mean-squared change of the density matrix elements and 10^{-4} a.u. for the maximum change of the density matrix elements. Nakai and co-workers[76] and Shaw and St-Amant[22] have noted that DIIS will cause SCF calculations to oscillate at the final stage of the convergence due to the slight errors introduced by the Div&Con approximation for assembling the density matrix (see Equation 2-17). In our HF Div&Con calculations, the DIIS technique was turned off when the root-mean-squared change of the density matrix elements reaches 10^{-4} a.u.. We also found that although the DIIS works in the early stage of the SCF procedure, we get the best performance in the SCF convergence when only two previous Fock matrices were stored in the DIIS calculations. One can see from Table 3-1 that the HF DC calculations usually requires more SCF cycles than the non-DC runs, however, for the polyglycine and polyalanine systems, the MFCC initial guess helps to reduce the number of SCF cycles in both DC and non-DC cases.

Table 3-1. Number of SCF cycles needed to reach convergence for the SAD and MFCC initial guess at the HF/6-31G* level.

	Div&Con		Non-Div&Con*	
System	SAD initial guess	MFCC initial guess	SAD initial guess	MFCC initial guess
Gly_6	18	10	12	7
Gly_{10}	18	11	12	7
Gly_{20}	18	10	12	6
Gly_{30}	18	10	12	6
Gly_{40}	18	8	12	7
α-$(Ala)_{10}$	18	15	12	9
α-$(Ala)_{20}$	16	12	12	9
α-$(Ala)_{30}$	16	12	12	8
α-$(Ala)_{40}$	15	12	12	8

*In the SCF procedure of non-Div&Con case, every 10 previous Fock matrices were stored in the DIIS calculations; while for the Div&Con case, every 2 previous Fock matrices were stored in the DIIS calculations until the root-mean-squared change of the density matrix elements reaches 10^{-4} a.u., after that, the DIIS technique was turned off.

3.3 Residue-centric Core Region versus Atom-centric Core Region

Previously, all the calculations used a residue based definition for the core region. We have also examined an atom based subsetting strategy for the core region in polyglycines and polyalanines. One can see from Table 3-2, the converged total energies using atom-centric core region were almost identical to those using a residue-based cutoff. Indeed, the overall mean unsigned deviation is as low as 0.054 kcal mol^{-1}. This is an attractive aspect of the divide and conquer approach since it allows for parallelization at the atom level rather than at the much larger residue based cutoff scheme.

Table 3-2. The converged total energies (a.u.) (at the HF/6-31G* level) using two different subsetting schemes: residue-based with buffer of 5Å and atom-based with a buffer of 7Å. (MUD: mean unsigned deviation)

System	Residue-centric core region	Atom-centric core region	Deviation (kcal mol^{-1})
Gly_{10}	-2314.783296	-2314.783272	-0.015
Gly_{20}	-4382.595749	-4382.595726	-0.014
Gly_{30}	-6450.407962	-6450.407938	-0.015
Gly_{40}	-8518.221662	-8518.221679	0.011
α-$(Ala)_{20}$	-5164.086850	-5164.086911	0.038
α-$(Ala)_{30}$	-7622.660188	-7622.660373	0.116
α-$(Ala)_{40}$	-10081.238571	-10081.238839	0.168
MUD			0.054

3.4 Validation on Three-dimensional Protein Systems

No previous studies have utilized the divide-and-conquer HF approach on three-dimensional globular proteins. In order to address this point, we have validated the accuracy of divide-and-conquer HF/6-31G* calculations on eleven real proteins. The systems ranged from 304 atoms to 608 atoms and are listed in Table 3-3. The proteins consisted of $\alpha-$helical structures (see Figure 3-8a) or are mixed $\alpha-\beta-$structures (see Figure 3-8b). As shown in Table 3-3, the largest deviation is 2.25 kcal mol^{-1} and the overall mean unsigned deviation is only 0.97 kcal mol^{-1} compared to standard full system calculations. Importantly, the observed error is large than what was observed for the one-dimensional examples, but is still within acceptable limits. This study sets the stage for the wide application of divide-and-conquer calculations on real protein systems. Furthermore, we have found that for five proteins, the divide-and-conquer HF calculations are not able to reach convergence using the simple SAD initial guess, while all the cases converged using the MFCC initial guess. Therefore, we conclude that the quality of initial guess plays an important role in insuring the convergence of divide-and-conquer calculations. Fragment-based electron density provides a much better quality initial guess with linear-scaling computational cost and, ultimately, much less computational time when compared to full system calculations.

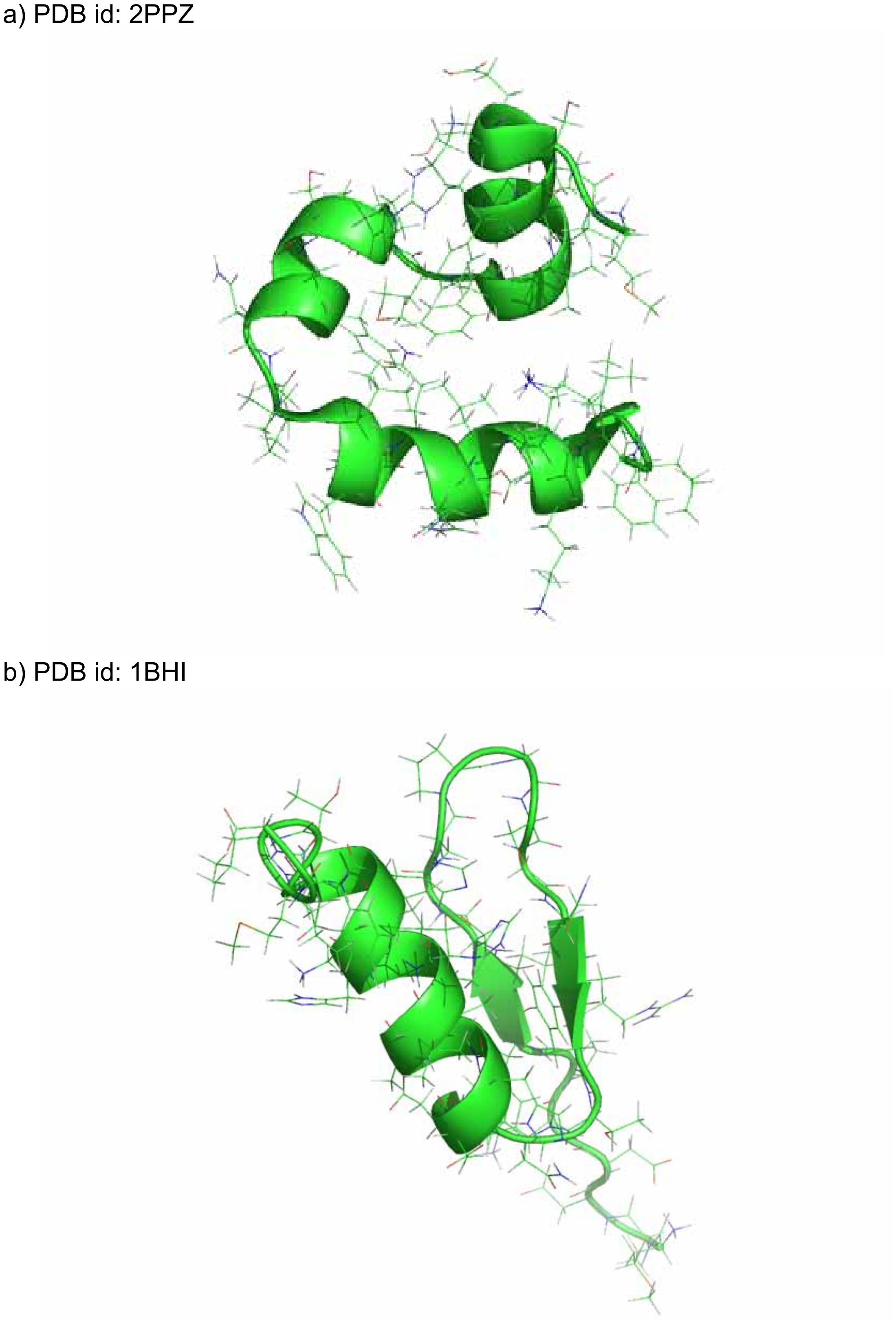
3.5 Conclusions

In this study, the divide-and-conquer HF theory was revisited in order to examine its potential to study three-dimensional constructs and a new and effective initial guess scheme was introduced. We first validated the accuracy of the divide-and-conquer HF/6-31G* calculations on eleven three-dimensional globular proteins. The overall mean unsigned error was within 1 kcal mol^{-1} when compared to standard full

Table 3-3. The total energies (a.u.) of three-dimensional globular proteins obtained using standard full system HF/6-31G* calculations and divide-and-conquer HF/6-31G* calculations using the MFCC initial guess. (MUD: mean unsigned deviation)

System	Number of atoms	Number of basis functions	Standard full system calculation	Div&Con using MFCC initial guess	Deviation (kcal mol^{-1})
Trp-cage	304	2610	-7439.721780	-7439.722124	-0.22
1VTP	396	3418	-10014.756053	-10014.755741*	0.20
1BBA	582	5033	-15103.299186	-15103.302595	-2.14
1AML	598	5178	-15140.895905	-15140.897305*	-0.88
1BHI	591	5124	-15989.697592	-15989.696544	0.66
1BZG	573	4851	-13680.602670	-13680.602916*	-0.15
2JPK	589	5000	-13854.809422	-13854.810188*	-0.48
2KCF	576	4991	-14599.178617	-14599.180118	-0.94
2PPZ	608	5111	-14957.602116	-14957.605696	-2.25
2RLK	588	5089	-14589.701015	-14589.702771*	-1.10
2YSC	578	5108	-14634.254517	-14634.257181	-1.67
MUD					0.97

* Did not converge using the SAD initial guess.



a) PDB id: 2PPZ

b) PDB id: 1BHI

Figure 3-8. Two representative three-dimensional protein structures studied in this thesis.

system calculations. Furthermore, we found that the fragment-based initial guess using the MFCC approach reduces the number of SCF cycles for polyglycine and polyalanine systems. Moreover, the MFCC initial guess facilitated SCF convergence for several of the globular proteins, where the SAD initial guess was unable to yield a converged wavefunction.

CHAPTER 4
PROTEIN NMR CHEMICAL SHIFT CALCULATIONS BASED ON THE AUTOMATED FRAGMENTATION QM/MM APPROACH

4.1 Introduction

NMR spectroscopy is a powerful tool used to study the three-dimensional structure and dynamics of biological systems.[83,84] Since the determination of NMR spectra does not require proteins to be crystallized, it can be applied to proteins in a variety of situations including the solid state and solution phases.[85] NMR chemical shifts accurately reflect the local chemical environment at atomic resolution. Thus, the secondary structure of proteins can be determined from NMR chemical shifts.[86] Recent studies show that in combination with traditional molecular mechanical force fields[87] or *de novo* protein structure sampling techniques[88-90], protein structures can be derived using ^{1}H, ^{13}C and ^{15}N NMR chemical shifts.

Several empirical models have been developed to compute NMR chemical shifts for proteins.[91,92] However, the success of the empirical methods requires a “basis set” of known chemical shifts to derive a set of well-tuned parameters. It is not a trivial process to generalize empirical approach to handle proteins with nonstandard residues, metal cofactors (as in metalloenzymes) and protein-ligand complexes. Linear-scaling semiempirical quantum mechanical NMR chemical shift computations have been reported by Wang and Merz, which generalize the computation of chemical shifts to many environments.[93,94] Much effort has also been devoted to make modern HF and DFT quantum mechanical calculations applicable to 100-200 atom NMR chemical shift calculations.[95-97] DFT and *ab initio* calculations clearly offer the most robust theoretical model for the prediction of NMR chemical shifts. However, it has not been practical to apply standard all-electron quantum chemistry methods to macromolecules because of the poor scaling of *ab initio* and DFT methods. The rate-limiting step in calculating the NMR chemical shieldings is the solution of the coupled perturbed Hartree-Fock (CPHF) equation. The transformation from atomic orbital (AO) two-electron integrals to molecular orbital (MO) two-electron integrals involved in solving the CPHF equation scales to the fifth power of the molecular size $O(M^5)$. Over the past two decades, advances in quantum chemistry have reduced the scaling to $O(M^3)$.[95,98] More recently, Kussmann and Ochsenfeld introduced a linear-scaling *ab initio* method for calculating NMR chemical shifts, and applied it to systems with over one thousand atoms.[99,100] Gao *et.al.* have also reported a fragment molecular orbital (FMO) method for NMR chemical shielding calculations at the HF level.[101]

In this chapter, we propose a more efficient automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM) which can be applied to routinely calculate the *ab initio* NMR chemical shieldings for proteins of any size. In our automated fragmentation approach, the entire protein is divided into individual fragments. Residues within a certain buffer region surrounding each fragment are included in the QM calculations to preserve the chemical environment of the divided fragment.[17,19,102,103] The remainder of the system outside the buffer regions is described by molecular mechanics. Each fragment-centric QM/MM calculation is carried out separately; hence, the method is trivially parallel. The many-body effects are intrinsically taken care of within the QM region, which is in contrast to the FMO implementation of the NMR chemical shift computation where only two-body interactions are taken into account.[101] The AF-QM/MM approach generates each fragment automatically and is applied to the entire protein system, not just to a small reaction center or region that is typical in standard QM/MM methods.[104,105]

The AF-QM/MM NMR approach is inspired by the fact that the NMR chemical shifts are local physical properties. For example, it has been shown that good accuracy can be achieved using the hybrid QM/MM method to calculate chemical shifts,[64,106,107] since the local electron density distribution around the atoms of interest is adequate to describe the QM effects on the NMR chemical shifts. The local short ranged interactions are largely contributed from the sequentially connected residues, hydrogen bonding, ring current effects and other van der Waals and electrostatic interactions from non-neighboring residues that are in close contact. By non-neighboring residues, we mean that the two residues are not sequentially connected to each other. In this approach, high-level *ab initio* methods can be applied to effectively describe the major interactions contributed to the NMR chemical shifts while the MM model gives the long-range electrostatic potential. To demonstrate the utility of the AF-QM/MM approach for linear-scaling *ab initio* NMR chemical shift calculations on macromolecules, we have applied this approach on a globular mini-protein in this chapter.

4.2 Computational Approach

4.2.1 NMR Chemical Shift Computation

In the framework of the gauge-including atomic orbitals (GIAO) approach[108], the field-dependent atomic basis functions[109] are used to ensure the gauge invariance of NMR chemical shifts.

$$\chi_A(\vec{B}) = \exp\left[-\frac{i}{2c}(\vec{B}\times\vec{R}_A)\cdot\vec{r}\right]\chi_A(0) \tag{4-1}$$

where $\chi_A(0)$ is a Gaussian atomic orbital, B is the external magnetic field, R_A and r are spatial vectors of the nucleus and electron, respectively. The NMR chemical shielding tensor components σ_{ab} is the second derivative of the electronic energy with respect to the external magnetic field B and the magnetic moment of the nucleus μ.[62]

$$\sigma_{ab} = \left.\frac{\partial^2 E}{\partial B_a \partial \mu_b}\right|_{\vec{B},\vec{\mu}=0} = \sum_{\mu\nu} P_{\mu\nu} \frac{\partial^2 h_{\mu\nu}}{\partial B_a \partial \mu_b} + \sum_{\mu\nu} \frac{\partial P_{\mu\nu}}{\partial B_a} \frac{\partial h_{\mu\nu}}{\partial \mu_b} \tag{4-2}$$

where $h_{\mu\nu}$ is the one-electron core Hamiltonian and $P_{\mu\nu}$ is the element of the density matrix. For closed-shell systems,

$$P_{\mu\nu} = \sum_{a}^{N_{occ}} C_{\mu a} C_{\nu a}^{*} \tag{4-3}$$

where $C_{\mu a}$ are the molecular orbital coefficients.

In the current version of the automated fragmentation approach, the MM atoms augment the one-electron Hamiltonian by adding the following QM/MM electrostatic interaction term

$$H_{QM/MM} = -\sum_{q} \frac{Z_q}{|R_q - r|} \tag{4-4}$$

where Z_q and R_q are the atomic charges and positions of the MM atoms. As shown by Cui and Karplus[64], the incorporation of MM atoms accounts for environmental effects on the chemical shielding tensors of the atoms in QM regions. The additional one-electron Hamiltonian $H_{QM/MM}$ perturbs both the density matrix and the first derivative of the density matrix resulting in contributions to both terms of Equation 4-2 for the chemical shielding tensor calculations.

The incorporation of point charges into the NMR chemical shielding calculations was handled by the Gaussian 03 program[110] without additional modification. AF-QM/MM approach makes it possible to carry out NMR shielding calculations on real protein systems with thousands of atoms.

4.2.2 MD Simulations

One of the Trp-cage NMR structures[111] was used as the starting point for the simulations. The protein was solvated in an octahedral TIP3P water box[112] with each side at least 8 Å from the nearest solute atom. The net charge of the entire system was neutralized by applying a uniform neutralized plasma[113,114]. The SHAKE algorithm[115] was employed to constrain X-H (X=C,N,O and S) bonds to their equilibrium values. The system was minimized and then gradually heated up from 0K to 300K with decreasing weak restraints on the heavy atoms of the protein. During the last step of equilibration, the restraints were removed entirely, and the production simulations were performed at 300K for 10ns with a 2 fs time step. Constant temperature was maintained using Berendsen's method[116] with a coupling strength of 1.0 ps. Snapshots for subsequent analysis were taken every 2 ps. All simulations were performed using the PMEMD module from the AMBER suite of programs[117].

4.3 Results and Discussion

The AF-QM/MM approach was used to compute the ^{1}H, ^{13}C and ^{15}N chemical shifts of the mini-protein Trp-cage, which is shown in Figure 4-1. Trp-cage has 20 residues and 304 atoms in total. The structure was taken from one of the NMR structures (pdb id:1L2Y[111]) and it was optimized using Sander from the AMBER program suite[117] in order to remove bad contacts prior to subsequent computations. All

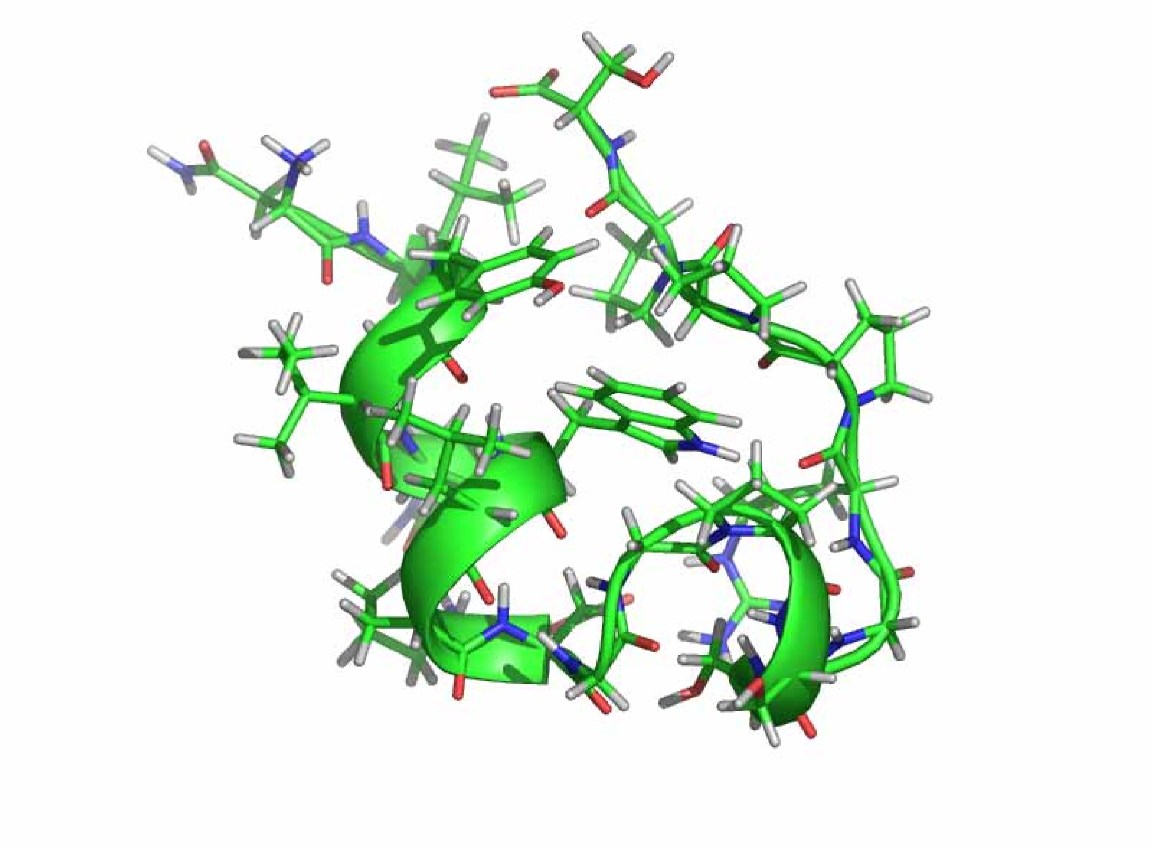


Figure 4-1. The NMR structure of Trp-cage (pdb entry: 1L2Y)

the *ab initio* NMR chemical shift calculations on the optimized conformation were carried out using the Gaussian 03 program.[110] In order to validate and select the best atomic point charge model to reproduce the electrostatic field in our NMR calculations, we evaluated several charge sets. From *ab initio* calculations we examined Mulliken charges[118] and NPA charges[119] from natural population analysis. These two charge models were derived from automated fragmentation calculations at the *ab initio* level. The subsetting scheme is the same as the one used for the NMR chemical shielding calculations, but we only perform QM calculations on each subsystem without the MM environment, because the point charges have not been determined at the initial stage. Only the atomic charges on the core region are extracted from each QM calculation. For comparison, CM1[120] and CM2[121] charges in conjuction with AM1[122] and PM3[123] methods were derived using the linear-scaling program DivCon[124]. By applying the divide-and-conquer algorithm[16,17,19], the computational expense of AM1 and PM3 calculations has been reduced to linear-scaling with a small prefactor due to the semi-empirical Hamiltonian; hence, the computation is much faster than the *ab initio* atomic charge calculations. Empirical point charges from AMBER force field were also used for comparison.

Table 4-1. Comparison of AF-QM/MM and full system HF/3-21G isotropic shielding constants for the ^{1}H, ^{13}C and ^{15}N atoms in Trp-cage.

Charge	^{1}H (ppm)			^{13}C (ppm)			^{15}N (ppm)		
model	MaxE	MUE	RMSE	MaxE	MUE	RMSE	MaxE	MUE	RMSE
No charges	0.84	0.17	0.22	2.08	0.41	0.59	5.07	1.77	2.21
AMBER	0.29	0.05	0.07	0.52	0.09	0.13	1.17	0.37	0.46
Mulliken	0.27	0.07	0.09	0.65	0.13	0.19	1.94	0.38	0.57
NPA	0.30	0.06	0.09	0.57	0.11	0.16	1.31	0.34	0.44
AM1/CM1	0.34	0.06	0.08	0.74	0.14	0.21	1.77	0.57	0.71
PM3/CM1	0.32	0.05	0.07	0.48	0.09	0.13	1.32	0.36	0.48
AM1/CM2	0.32	0.05	0.07	0.57	0.11	0.16	1.49	0.44	0.57
PM3/CM2	0.32	0.05	0.07	0.57	0.10	0.15	1.60	0.42	0.56

MaxE: maximum error; MUE: mean unsigned error; RMSE: root mean squared error.

We first carried out AF-QM/MM calculations using HF/3-21G for each QM part and different charge models for the MM environment. Table 4-1 shows the quality of NMR isotropic shielding constants for the Trp-cage based on the AF-QM/MM approach compared to a full system calculation. One can see from Table 4-1 that if MM charges are not employed, the root mean square errors (RMSEs) of chemical shieldings for ^{1}H, ^{13}C and ^{15}N are 0.22, 0.59 and 2.21 ppm, respectively. In contrast, the AF-QM/MM approach with all charge models employed results in good agreement with full system calculations. The RMSEs for ^{1}H, ^{13}C and ^{15}N shieldings are equal to or less than 0.09ppm, 0.21ppm, and 0.71ppm, respectively. The incorporation of MM point charges reduces the RMSEs, by ~2.5-5 fold for all the charge models, and also reduces, by

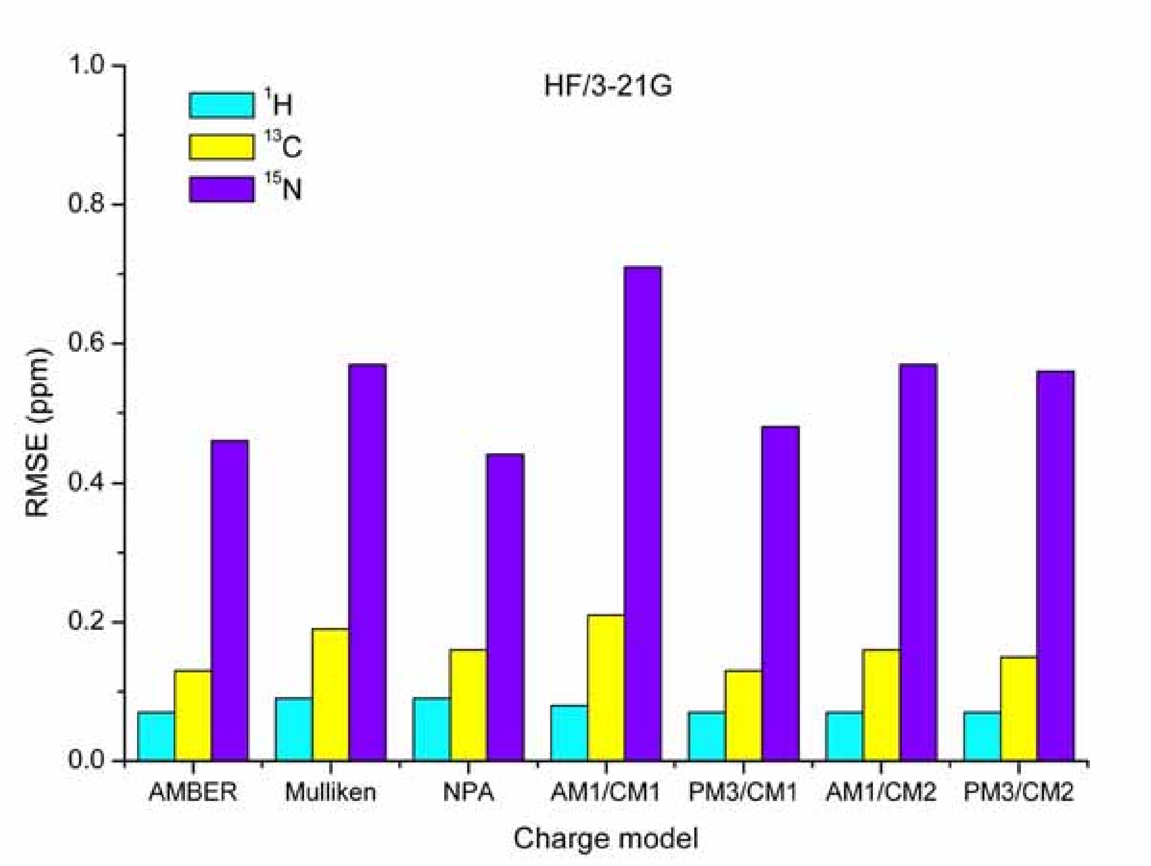


Figure 4-2. The RMSEs of the ^{1}H, ^{13}C and ^{15}N chemical shielding constants using the AF-QM/MM approach with different point charge models at the HF/3-21G level (compared to conventional full system calculations)

about the same order of magnitude, the maximum deviation and the mean deviation. As expected, the electrostatic potential of the MM environment on the NMR chemical shieldings is important and cannot be neglected. As shown in Figure 4-2, AM1/CM1 and Mulliken charges were the worst charge models to use in HF/3-21G NMR calculations. NPA works slightly better on the ^{15}N NMR chemical shieldings than CM1 and CM2 charges derived from semiempirical AM1 and PM3 calculations, and has about the same accuracy for the ^{13}C shieldings. But for ^{1}H, the CM1 and CM2 charges outperforms the NPA charges obtained from relatively expensive *ab initio* calculations. Interestingly, the point charge model from the AMBER force field gives the lowest

RMSEs of 0.07ppm and 0.13ppm for ^{1}H and ^{13}C shieldings, respectively. These charges also give a RMSE of 0.46ppm, which is similar to the lowest observed RMSE of 0.44ppm using the NPA charge model for ^{15}N chemical shieldings. Hence, we conclude that for Trp-cage, the AMBER point charge model is the best for the AF-QM/MM approach at the HF/3-21G level.

Table 4-2. Comparison of AF-QM/MM and full system HF/6-31G** isotropic shielding constants for the ^{1}H, ^{13}C and ^{15}N atoms in Trp-cage.

Charge	^{1}H (ppm)			^{13}C (ppm)			^{15}N (ppm)		
model	MaxE	MUE	RMSE	MaxE	MUE	RMSE	MaxE	MUE	RMSE
No charges	0.93	0.15	0.21	2.31	0.43	0.63	4.87	1.77	2.21
AMBER	0.24	0.05	0.07	0.45	0.08	0.12	0.89	0.36	0.43
Mulliken	0.23	0.05	0.06	0.61	0.10	0.15	1.27	0.28	0.40
NPA	0.35	0.06	0.08	0.85	0.13	0.20	0.92	0.37	0.45
AM1/CM1	0.28	0.05	0.07	0.78	0.14	0.21	1.54	0.54	0.67
PM3/CM1	0.28	0.04	0.06	0.42	0.08	0.11	1.07	0.31	0.42
AM1/CM2	0.27	0.05	0.06	0.51	0.10	0.15	1.25	0.40	0.51
PM3/CM2	0.28	0.05	0.06	0.49	0.09	0.14	1.35	0.36	0.49
PM3/CM2*	0.29	0.04	0.06	0.56	0.10	0.15	1.34	0.40	0.53

MaxE: maximum error; MUE: mean unsigned error; RMSE: root mean squared error.
*Calculations were carried out using HF/6-311G**

We also performed HF/6-31G** calculations to further explore the basis set dependence and these results are summarized in Table 4-2. As was found for the HF/3-21G calculations, the incorporation of a point charge model is superior to calculations without the MM environment. Employing any of the charge models described herein, the RMSEs for the ^{1}H, ^{13}C and ^{15}N NMR chemical shieldings are equal to or less than 0.08ppm, 0.21ppm, and 0.67ppm, respectively. Among all the charge models, AM1/CM1 has the largest RMSEs for ^{13}C and ^{15}N as shown in Figure 4-3, while Mulliken charges work well with HF/6-31G**, which is not consistent with what was observed in the previous HF/3-21G calculations. The AMBER, Mulliken, NPA, PM3/CM1, AM1/CM2 and PM3/CM2 charge models are all very similar; however, the NPA charge model has the largest RMSEs for the ^{1}H and ^{13}C NMR chemical shieldings among this set. Finally, we carried out HF calculations using the valence triple zeta 6-311G** basis set using the PM3/CM2 charge model. The accuracy obtained is very close to calculations using the 6-31G** basis set as summarized in the Table 4-2.

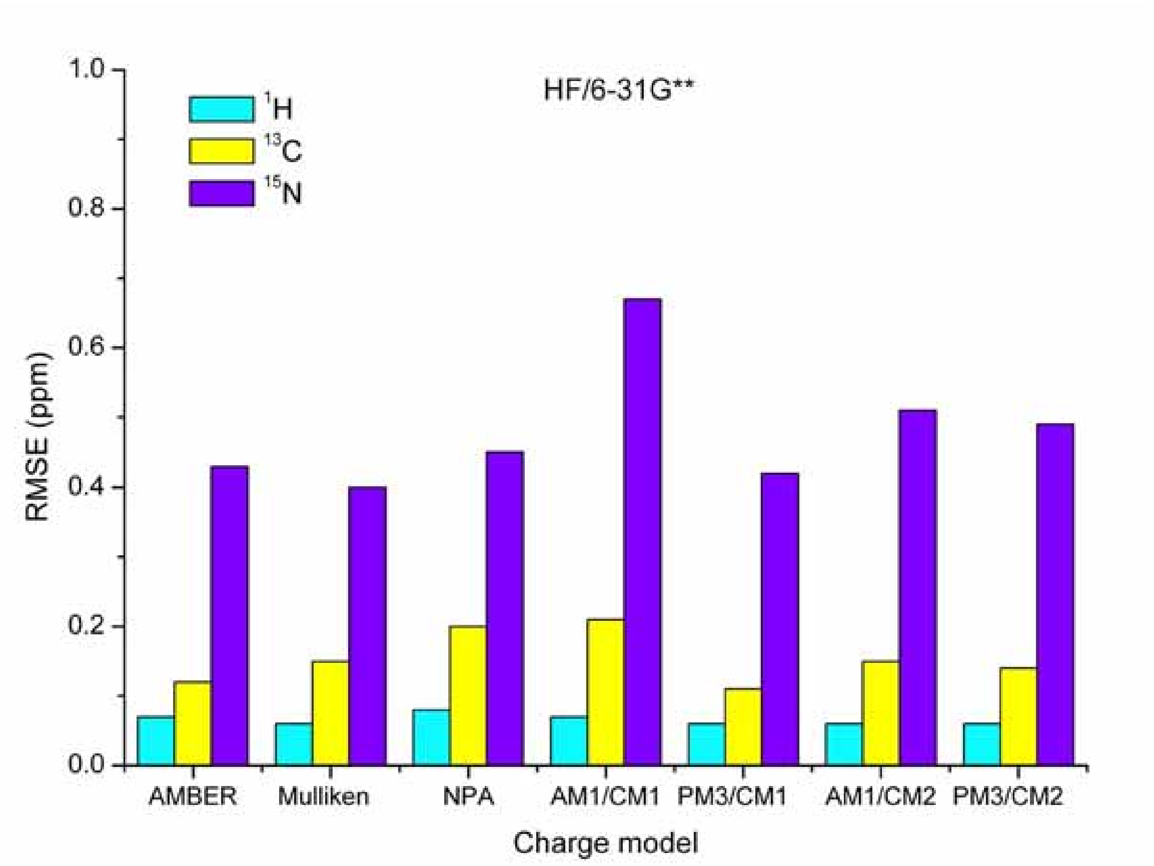


Figure 4-3. The RMSEs of the ^{1}H, ^{13}C and ^{15}N chemical shielding constants using the AF-QM/MM approach with different point charge models at the HF/6-31G** level (compared to conventional full system calculations)

The AF-QM/MM approach using DFT theory was also performed using B3LYP/6-31G**. Again, as indicated in Table 4-3, the point charge model should be included in the AF-QM/MM NMR shielding calculations using DFT. The RMSEs for the ^{1}H, ^{13}C and ^{15}N NMR chemical shieldings using all the charge models are equal to or less than

Table 4-3. Comparison of AF-QM/MM and full system B3LYP/6-31G** isotropic shielding constants for the ^{1}H, ^{13}C and ^{15}N atoms in Trp-cage.

Charge	^{1}H (ppm)			^{13}C (ppm)			^{15}N (ppm)		
model	MaxE	MUE	RMSE	MaxE	MUE	RMSE	MaxE	MUE	RMSE
No charges	1.04	0.15	0.21	2.70	0.52	0.75	5.20	1.99	2.46
AMBER	0.27	0.05	0.07	0.88	0.17	0.22	1.03	0.45	0.53
Mulliken	0.32	0.06	0.08	0.93	0.23	0.30	2.26	0.57	0.78
NPA	0.28	0.06	0.08	1.12	0.24	0.32	1.21	0.40	0.50
AM1/CM1	0.20	0.05	0.07	0.84	0.20	0.27	1.50	0.57	0.71
PM3/CM1	0.20	0.05	0.06	0.92	0.16	0.21	1.00	0.37	0.46
AM1/CM2	0.19	0.05	0.06	0.88	0.17	0.22	1.18	0.44	0.55
PM3/CM2	0.22	0.05	0.06	0.90	0.17	0.22	1.31	0.42	0.55
PM3/CM2*	0.23	0.05	0.06	0.67	0.20	0.26	1.75	0.63	0.75

MaxE: maximum error; MUE: mean unsigned error; RMSE: root mean squared error.
*Calculations were carried out using B3LYP/6-311G**

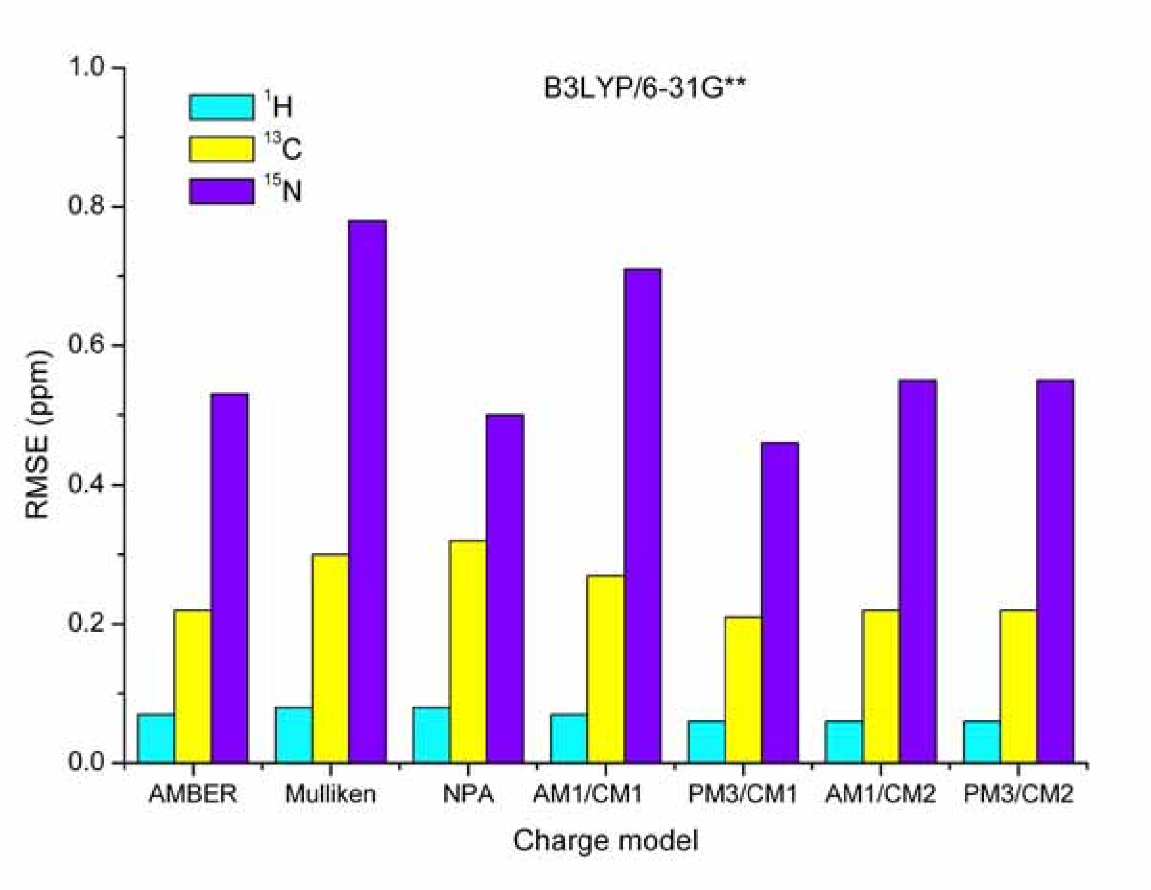


Figure 4-4. The RMSEs of the ^{1}H, ^{13}C and ^{15}N chemical shielding constants using the AF-QM/MM approach with different point charge models at the B3LYP/6-31G** level (compared to conventional full system calculations)

Table 4-4. Comparison of ^{1}H NMR chemical shifts between the AF-QM/MM *ab initio* calculations and experiment for Trp-cage.[a]

Method	R^2	RMSE (ppm)[b]	MUE (ppm)[b]	MSE (ppm)[b]	Reference (ppm)
HF/3-21G	0.9237	0.51	0.41	0.15	33.8271
HF/6-31G**	0.9396	0.47	0.37	0.00	32.3347
HF/6-311G**	0.9394	0.47	0.37	0.01	32.4722
B3LYP/6-31G**	0.9541	0.39	0.30	-0.10	31.7461
B3LYP/6-311G**	0.9551	0.38	0.29	-0.03	31.9949
B3LYP/6-31+G*	0.9450	0.44	0.34	-0.18	32.0579
B3LYP/6-31+G**	0.9473	0.43	0.34	-0.11	31.6414
B3LYP/6-311++G**	0.9512	0.40	0.32	-0.05	31.9006
B3LYP/6-31G** (average)	0.9696	0.34	0.27	-0.13	31.7461
MNDO/NMR[93]	0.8897	0.49	0.40	0.01	41.2062
SHIFTX[91] [c]	0.9440	0.27	0.18	-0.01	
SHIFTS[92]	0.9800	0.24	0.17	-0.01	

a) PM3/CM2 was used to generate the point charges for the MM environment. Values are referenced to the ^{1}H isotropic shielding constants computed for TMS in the gas phase at each *ab initio* level.
b) MaxE: maximum error; MUE: mean unsigned error; RMSE: root mean squared error.
c) SHIFTX does not calculate the NMR chemical shifts of ^{1}H on the aromatic rings of Tyr3 and Trp6.

0.08ppm, 0.32ppm, and 0.78ppm, respectively, similar to what was observed using HF/6-31G** calculations. Furthermore, Mulliken charges and NPA charges give larger RMSEs on ^{1}H and ^{13}C than other charge models, and Mulliken charges gave the largest RMSE of 0.78 ppm on ^{15}N among all the charge models. AMBER, PM3/CM1, AM1/CM2 and PM3/CM2 perform similarly with AM1/CM1 being the worst among all of the empirical charge models. B3LYP using the 6-311G** basis set combined with PM3/CM2 charge model yields similar MUEs and RMSEs for the ^{1}H and ^{13}C NMR chemical shieldings when compared to the 6-31G** basis set. Although the RMSE for B3LYP/6-311G** increases from 0.55 ppm (using B3LYP/6-31G**) to 0.75ppm on the ^{15}N shieldings, it still gives acceptable agreement with full system calculations.

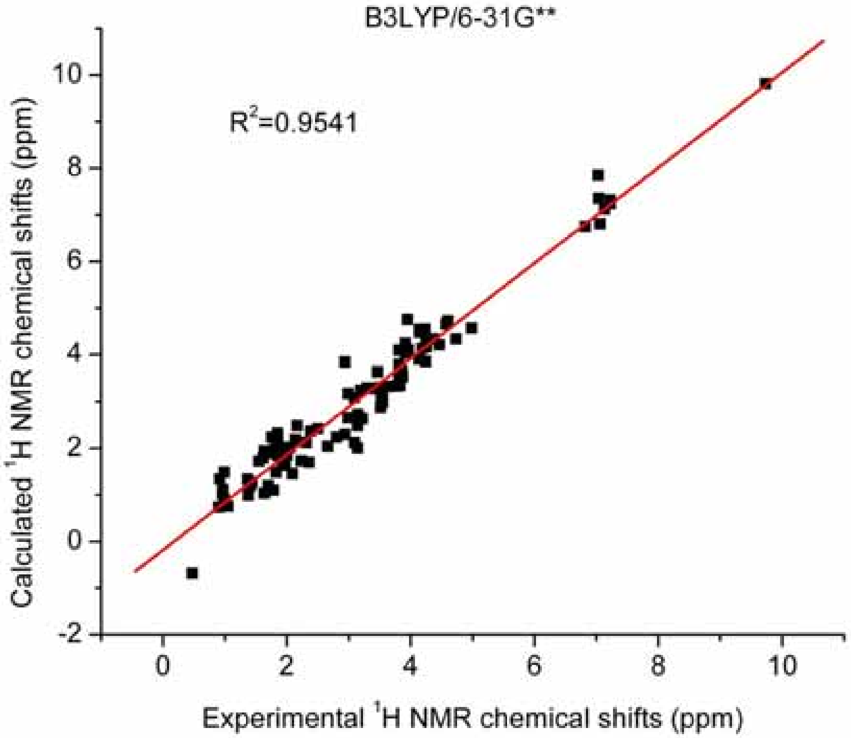


Figure 4-5. The correlation between experimental ^{1}H NMR chemical shifts (excluding the exchangeable protons) and calculated chemical shifts using the AF-QM/MM approach. The QM calculations were done at the B3LYP/6-31G** level. The PM3/CM2 charge model was used to derive the MM point charges.

We also analyzed the correlation and deviation of our computed ^{1}H NMR chemical shifts (exchangeable protons are excluded) with experimental data for the Trp-cage (see Table 4-4). PM3/CM2 was used to generate the point charge for the MM environment since it is one of the best polarizable charge models for NMR chemical shift calculations as observed previously for HF/6-31G** and B3LYP/6-31G** calculations (the results for other charge models are listed in Tables 4-5, 4-6 and 4-7). One can see from Table 4-4, DFT gives better correlation with experimental observations than HF calculations. Among all the DFT calculations, B3LYP with the 6-31G**, 6-311G** and 6-311++G** basis sets have smaller RMSEs (between 0.38ppm and 0.40ppm) than other basis sets and their correlation with experimental ^{1}H NMR chemical shifts are 0.9541, 0.9551 and 0.9512, respectively. Figure 4-5 and 4-6 show

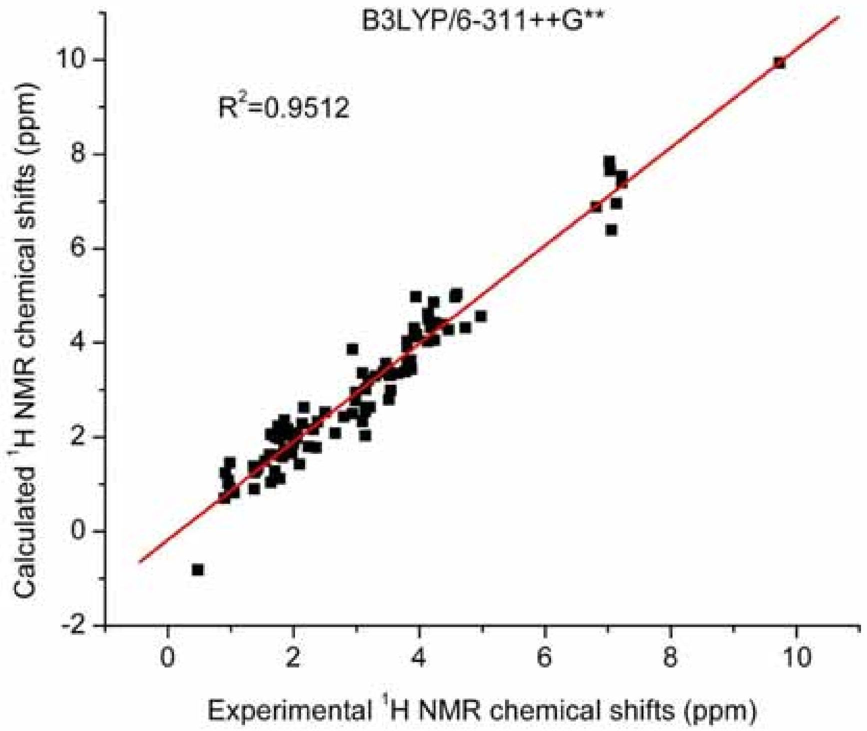


Figure 4-6. The correlation between experimental ^{1}H NMR chemical shifts (excluding the exchangeable protons) and calculated chemical shifts using the AF-QM/MM approach. The QM calculations were done at the B3LYP/6-311++G** level. The PM3/CM2 charge model was used to derive the MM point charges.

Table 4-5. Comparison of ^{1}H NMR chemical shifts between the AF-QM/MM HF/3-21G calculations and experiment for Trp-cage.[a]

Method	R^2	RMSE(ppm)[b]	MUE(ppm)[b]	MSE(ppm)[b]	Reference (ppm)
No charges	0.9110	0.55	0.45	0.14	33.8271
AMBER	0.9235	0.51	0.41	0.14	
Mulliken	0.9246	0.50	0.41	0.14	
NPA	0.9251	0.50	0.41	0.14	
AM1/CM1	0.9226	0.51	0.41	0.14	
PM3/CM1	0.9237	0.51	0.41	0.15	
AM1/CM2	0.9237	0.51	0.41	0.14	
PM3/CM2	0.9237	0.51	0.41	0.15	

a) Values are referenced to the ^{1}H isotropic shielding constant computed for TMS in the gas phase at the HF/3-21G level.

b) MaxE: maximum error; MUE: mean unsigned error; RMSE: root mean squared error.

the correlations between experimental ^{1}H NMR chemical shifts and calculated chemical shifts using B3LYP/6-31G** and B3LYP/6-311++G** for each QM region, respectively.

Table 4-6. Similar to Table 4-5, but for HF/6-31G** calculations.

Method	R^2	RMSE (ppm) [b]	MUE (ppm) [b]	MSE (ppm) [b]	Reference (ppm)
No charges	0.9313	0.50	0.41	0.00	32.3347
AMBER	0.9395	0.47	0.37	0.00	
Mulliken	0.9403	0.47	0.37	0.00	
NPA	0.9407	0.47	0.36	0.00	
AM1/CM1	0.9386	0.47	0.38	0.00	
PM3/CM1	0.9395	0.47	0.37	0.00	
AM1/CM2	0.9394	0.47	0.37	0.00	
PM3/CM2	0.9396	0.47	0.37	0.00	

Table 4-7. Similar to Table 4-5, but for B3LYP/6-31G** calculations.

Method	R^2	RMSE (ppm) [b]	MUE (ppm) [b]	MSE (ppm) [b]	Reference (ppm)
No charges	0.9477	0.42	0.33	-0.10	31.7461
AMBER	0.9546	0.39	0.30	-0.10	
Mulliken	0.9545	0.39	0.30	-0.11	
NPA	0.9540	0.39	0.31	-0.11	
AM1/CM1	0.9536	0.40	0.31	-0.11	
PM3/CM1	0.9539	0.39	0.30	-0.10	
AM1/CM2	0.9541	0.39	0.30	-0.11	
PM3/CM2	0.9541	0.39	0.30	-0.10	

Table 4-8. Comparison of ^{1}H NMR chemical shifts between the AF-QM/MM B3LYP/6-31G** calculations and experiment for Trp-cage using different buffer radii. (A: the buffer radius between any atom in the core region and the other atom outside the core region and at least one of the two atoms is a non-hydrogen atom; B: the buffer radius between one hydrogen atom in the core region and the other hydrogen atom outside the core region; C: the buffer radius between any atom in the core region and any heavy atom on an aromatic ring outside the core region.) The PM3/CM2 charge model was used to derive the MM point charges.

	Buffer radii (Å)		R^2	RMSE(ppm)	MUE(ppm)
A=4.0	B=3.0	C=5.0	0.9541	0.39	0.30
A=2.0	B=1.5	C=2.5	0.9104	0.52	0.36

To investigate the ring current effects on the ^{1}H chemical shifts, we focus on four protons: Gly11 Hα 2, Pro18 Hα , Hβ 2 and Hβ 3, which are close to Trp6. (see Figure

4-7a) Based on our fragmentation criterion, Trp6 is included in the QM region for the NMR chemical shift calculations on Gly11 and Pro18. As shown in Table 4-9, these four proton chemical shifts are highly perturbed when compared to AF-QM/MM calculations excluding Trp6 in the QM region. Furthermore, the AF-QM/MM results for these four proton chemical shifts agree with the full system all-electron calculations. It clearly shows that the AF-QM/MM approach captures the ring current effect experienced by NMR chemical shifts.

Since regions of proteins can have substantial conformational freedom, the discrepancy between computed and experimental chemical shifts may arise from the neglect of conformational sampling.[107,125-128] To take conformational fluctuations into account, we have chosen five NMR structures and performed AF-QM/MM calculations on each NMR structure using B3LYP/6-31G** and computed the average ^{1}H NMR chemical shifts over the five NMR structures. The average chemical shift of Pro18 Hβ3 becomes -0.20 ppm. The previous deviation of Pro18 Hβ3 from experiment is reduced from 1.15 ppm to 0.67 ppm as shown in Figure 4-8. In addition, the overall correlation for all the proton chemical shifts was increased from 0.9541 to 0.9696. Interestingly, both SHIFTS and SHIFTX have smaller RMSEs and MUEs between predicted and experimental shifts than does the *ab initio* calculations. Note that all the AF-QM/MM

Table 4-9. Experimental and theoretical predictions based on the AF-QM/MM approach[a] and conventional full system calculations[b] on four selected protons near Trp6.

Position	AF-QM/MM excluding Trp6 (ppm)	AF-QM/MM (ppm)	Full system (ppm)	Exp (ppm)
Gly11 Hα2	3.31	0.76	0.69	1.05
Pro18 Hα	4.60	2.05	2.02	2.66
Pro18 Hβ2	2.27	1.34	1.25	1.37
Pro18 Hβ3	2.04	-0.68	-0.75	0.47

a) B3LYP/6-31G** with the PM3/CM2 charge model
b) B3LYP/6-31G**

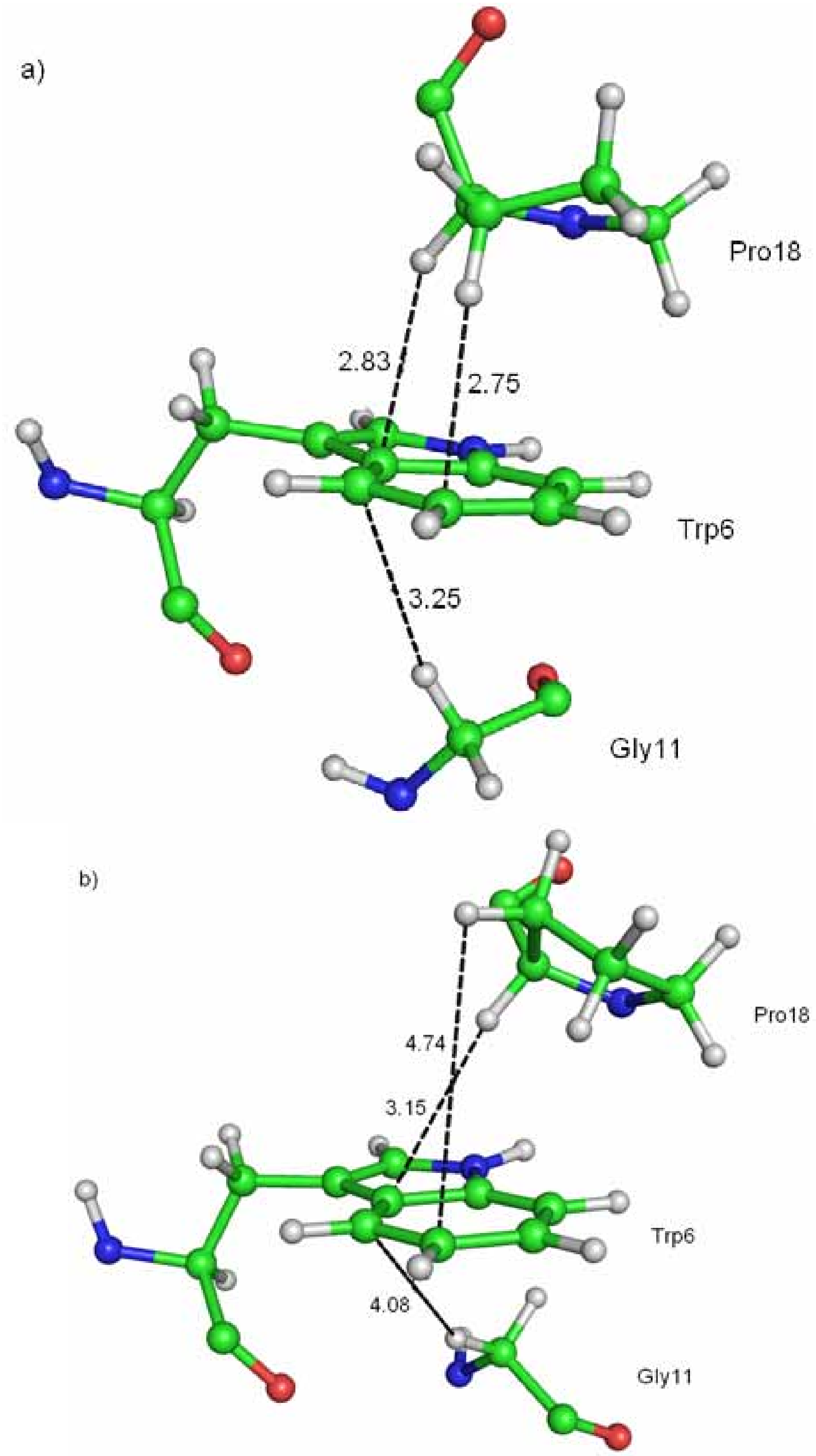


Figure 4-7. Structural details of the relative positions of Trp6, Gly11 and Pro18. (in Å) (a) One example of Pro18 in the down pucker conformation. (b) One representative configuration of a MD structure where Pro18 is in the up pucker conformation.

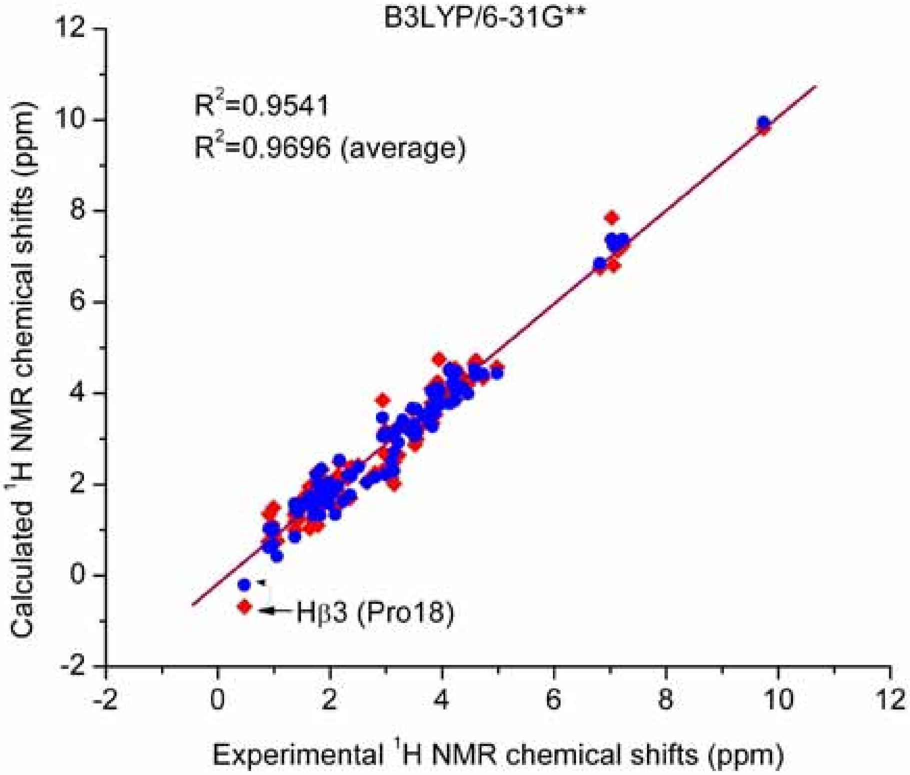


Figure 4-8. Unaveraged (red circles; see Figure 4-5) and the calculated average ^{1}H NMR chemical shifts based on 5 NMR structures using the AF-QM/MM approach (blue circles; B3LYP/6-31G** with PM3/CM2 charges).

calculations were performed in vacuum and that solvation effects are likely to be relevant. Furthermore, the choice of basis set and density functionals should be further explored to improve the accuracy of *ab initio* predictions.

Based on Nuclear Overhauser Effect (NOE) restraints, Pro18 was assigned as having the "down pucker" conformation in the NMR structures. Simmerling and co-workers have found both the down and up pucker conformations are populated during molecular dynamics simulations. Moreover, their empirical calculations using SHIFTS[92] reported a Hβ3 shift of -0.22 ppm on representative structures for Pro18 in the down pucker conformation,[129] which we predict to be -0.20 ppm on average using the AF-

Table 4-10. Average chemical shifts of Pro18 Hβ3 for the up and down pucker conformations. The average is based on a ratio of 33:67 between the up and down pucker conformations.

Method	Average of 5 up puckers (ppm)	Average of 5 down puckers (ppm)	Ensemble average (ppm)	Exp (ppm)
AF-QM/MM (B3LYP/6-31G**)	1.36	-0.20	0.31	0.47
SHIFTX	1.45	0.49	0.81	
SHIFTS	1.47	-0.09	0.42	

QM/MM approach. The measured ^{1}H NMR chemical shifts for Trp-cage were conformationally averaged in solution, which we further explored using molecular dynamics (MD) simulations employing the AMBER force field ff99sb[130]. Figure 4-9 shows the pseudorotation angles[131] for the five-member ring of Pro18 along the MD trajectory. The two bands in Figure 4-9 clearly indicate that there are two conformations (down and up pucker) for Pro18. We found that the populations for the down and up pucker conformations are 67% and 33%, respectively. For the up pucker conformation, the Hβ3 of Pro18 is shifted less upfield because the hydrogen moves away from Trp6 as clearly shown in Figure 4-7b relative to the down pucker conformation in Figure 4-7a. We also carried out AF-QM/MM calculations on 5 selected up pucker conformations from the MD trajectory using B3LYP/6-31G**. Table 4-10 shows the average Pro18 Hβ3 chemical shifts for 5 up and 5 down pucker conformations. The AF-QM/MM and SHIFTS predictions are consistent with each other. However, SHIFTX gives a prediction of 0.49 ppm on average for the down pucker conformation. Based on the ratio of 67:33 between the down and up pucker conformations, we obtain averaged chemical shifts for Pro18 Hβ3, which are listed in Table 4-10. The AF-QM/MM (0.31ppm) and SHIFTS (0.42ppm) predictions both agree with the experimental value of 0.47 ppm. SHIFTX overestimates this quantity due to overshooting the proton chemical shift for the down pucker conformation.

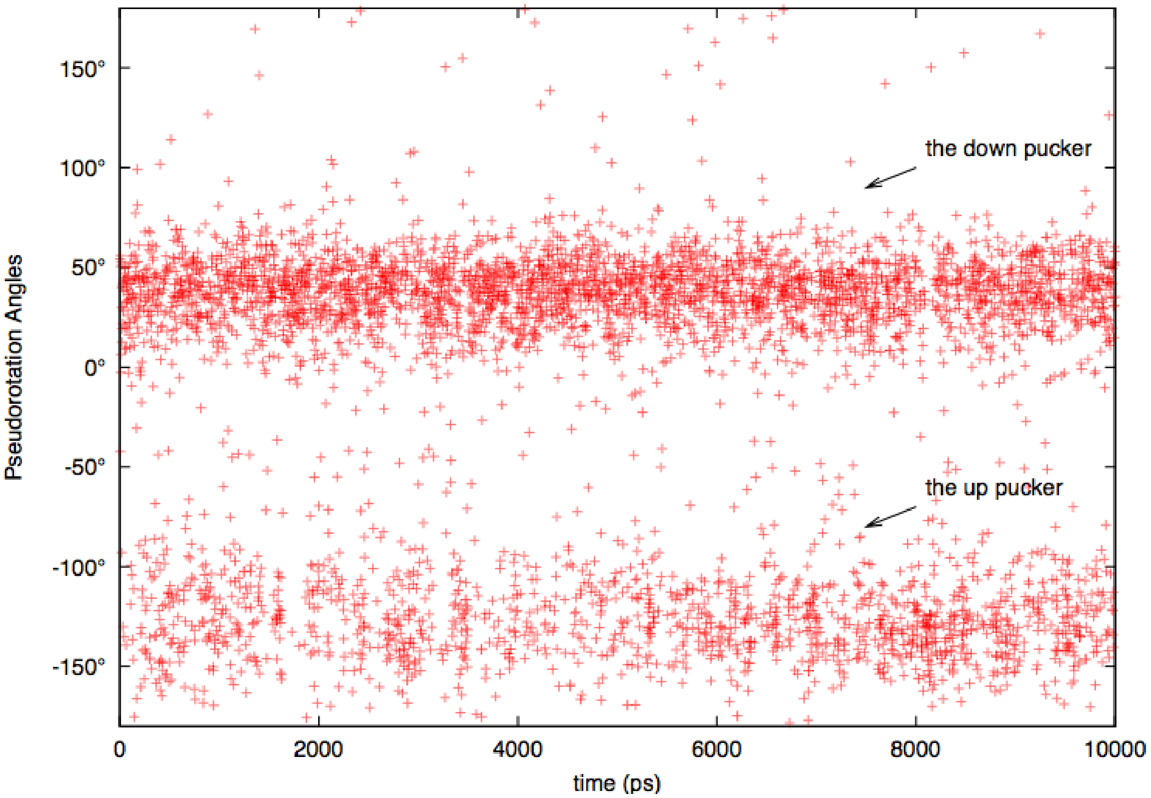


Figure 4-9. The pseudorotation angles of the five-member ring of Pro18 during the molecular dynamics simulations using PMEMD from the AMBER program suite.

4.4 Conclusions

The AF-QM/MM approach synthesizes quantum mechanics with molecular mechanics in order to study properties of protein system. It differs from the conventional QM/MM method in which only a part of the protein system is treated by quantum mechanics. In the AF-QM/MM approach, each residue along with its neighboring residues and non-neighboring residues that are spatially in close contact is computed by quantum mechanics, while all the long-range electrostatic interactions between distant non-neighboring residues are treated by molecular mechanics. The focus of this was not to compute a total energy (which it certainly could), but to focus on property computation. The goal of this study focussed on NMR chemical shieldings computed at the *ab initio* or DFT level using medium to large basis sets.

(1) The AF-QM/MM approach is computationally efficient and linear scaling. It combines the accuracy of quantum mechanics and the efficiency of molecular mechanics. Every automatically generated fragment normally contains less than 250 atoms consisting of C, H, O, N, and S. All the individual QM/MM calculations can be carried out at the HF and DFT level in parallel.

(2) The results from AF-QM/MM approach gave good agreements with conventional QM calculations on the entire protein. Indeed, the RMSEs for ^{1}H, ^{13}C and ^{15}N NMR chemical shieldings are equal or less than 0.09ppm, 0.32ppm, and 0.78ppm, respectively for all the HF and DFT (B3LYP) calculations described in this study.

(3) The electrostatic potential of the MM environment is important for NMR chemical shift calculations on the QM region. In general, we found that the AM1/CM1 charge model is not a good model for AF-QM/MM NMR chemical shielding calculations. Mulliken and NPA charges worked reasonably well at the HF/6-31G** level, but were worse than the empirical or semiempirical charge models using B3LYP/6-31G**. ^{15}N NMR chemical shieldings were found to be a special case where the NPA charge model was the second best. The AMBER, AM1/CM2, PM3/CM1 and PM3/CM2 charge modes performed similarly and work well with both the HF/6-31G** and B3LYP/6-31G** levels of theory. Importantly, the polarizable point charge models of AM1/CM2, PM3/CM1 and PM3/CM2 can be derived with much lower computational cost compared to *ab initio* atomic charge calculations.

(4) The correlations between experimental ^{1}H NMR chemical shifts and theoretical calculations are >0.95 for AF-QM/MM B3LYP calculations using the 6-31G**, 6-311G** and 6-311++G** basis sets. Averaging over five NMR structures increased the correlation between experiment and theory. The inclusion of conformational effects was found to be necessary to accurately predict NMR chemical shifts, which are sensitive to the local chemical environment.

Since the AF-QM/MM approach is trivially parallel, one can also inform protein structure and protein-ligand NMR based structure refinement utilizing *ab initio* NMR chemical shift calculations. Furthermore, the inclusion of solvation effects into the current model and other interesting applications based on the AF-QM/MM approach are ongoing in our laboratory.

CHAPTER 5
THE IMPORTANCE OF DISPERSION AND ELECTRON CORRELATION IN AB INITIO PROTEIN FOLDING

5.1 Introduction

The search for an energy-based “scoring” function that can routinely discriminate natively folded proteins from the non-native conformations is a major challenge for computational structural biology.[25] Based on the thermodynamic hypothesis, which states that the native state has the lowest free energy relative to misfolded states[26], current effort focuses on looking for reliable physics-based potentials that can distinguish native states from non-native ones.[27-31] Importantly, the free energy of the folded state in a protein is only 5-15Kcal/mol less than the denatured state ensemble.[32,33]; hence, it is clear that the final solution to this problem will require very high accuracy.

Not only is hydrogen bonding interactions important, but other non-covalent interactions, such as long range electrostatic and van der Waals interactions are important in defining protein structure. Recent theoretical and experimental studies have demonstrated the importance of non-covalent interactions.[132] In the protein folding process, the hydrophobic forces associated with non-polar residues results in the formation of the so-called hydrophobic core.[133,134] Indeed, a rather large attractive energy arises from the dispersion-dominated hydrophobic core collapse. By performing correlated *ab initio* calculations, Vondrasek *et al*. predicted the presence of a strong attraction inside the hydrophobic core of a small globular protein, which arises from the London dispersion energy between hydrophobic residues.[32] Riley and Merz, however, demonstrated that the extent of this interaction energy is mitigated by solvation effects reinforcing the well-known importance of solvation on the modeling of intramolecular interactions in proteins.[135] Moreover, studies by Fedorov *et.al*. showed the importance of dispersion in ligand-drug binding systems using *ab initio* MP2 calculation.[136] They also illustrated that the gas phase binding energy gap between the strongest binder and the weakest binder is much larger than the gap in the experimental binding free energies unless solvation effects are included. Therefore, it is clear that accurate solvation energies should be included in any effective energy-based scoring function for protein structure prediction.

Individual dispersion interactions are generally quite small, but when summed over all possible non-covalent interactions present in a protein the individual energies accumulate resulting in a significant contribution to the total free energy. To achieve accurate dispersion energies, correlated *ab initio* methods are required. Neither Hartree-Fock (HF) nor Density Functional Thoery (DFT) are formally able to capture these dispersion interactions.[32] Among all conventional *ab initio* electron correlation methods, second-order Møller-Plesset perturbation (MP2) theory is the least expensive non-empirical approach.

In the framework of Møller-Plesset (MP) perturbation theory, the electron correlation energy is the sum of second, third, fourth and higher order electron correlation energy:

$$\Delta E_{corr} = \Delta E^{(2)} + \Delta E^{(3)} + \Delta E^{(4)} + \cdots \tag{5-1}$$

MP2 which only takes the second-order correlation contribution into account generally gives a good estimate of the correlation energy.[137] In practice, MP2 is widely used as a benchmark calculation to describe the van der Waals interaction in dispersion-dominated complexes.[137-139] However, the second-order correlation energy obtained using MP2 ($\Delta E^{(2)}$) is not exactly equal to the dispersion energy. As has been shown by Cybulski *et.al.*[140] and Chalasinski and Szczesniak[141], $\Delta E^{(2)}$ can be decomposed into the intermolecular dispersion energy $\varepsilon_{disp}^{(20)}$, intramolecular electron correlation of the electrostatic energy $\varepsilon_{el}^{(12)}$, exchange correlation ε_{ex} and deformation correlation ε_{deform}.

$$\Delta E^{(2)} = \varepsilon_{el}^{(12)} + \varepsilon_{disp}^{(20)} + \varepsilon_{ex} + \varepsilon_{deform} \tag{5-2}$$

Although the dispersion energy frequently dominates the $\Delta E^{(2)}$ correlation energy, the intramolecular electron correlation and exchange correlation effects can have the same magnitude as the dispersion energy in some cases.[142] Hence, one needs to keep in mind that employing MP2 calculations to study biological systems not only captures the dispersion energy, but also includes local electron correlation and exchange effects.

Until recently, due to the relatively large size of proteins, it was not practical to apply standard all-electron quantum chemistry methods to compute the total energy of biomacromolecules because of the poor scaling of *ab initio* methods.[4] Much effort has been devoted to the development of linear-scaling methods over the past decades to compute the total energy of large molecular systems at the Hartree-Fock (HF) or density functional method (DFT) level.[6,9,12,16,17,42,68,69] The biggest challenge is to assemble the Fock matrix elements, which results in poor scaling properties due to long range Coulomb interactions. Fast multipole based approaches have successfully reduced the scaling in system size to linear[8,12,42,66,67] and made HF and DFT calculations affordable for larger systems when small to moderate sized basis sets are utilized. The more recently developed Fourier Transform Coulomb method of Fusti and Pulay[70,71] reduced the steep O(N^4) scaling in basis set size to quadratic and makes the calculations much more affordable with larger basis sets.[72] There is also a class of fragment-based methods for quantum calculation of protein systems including the divide and conquer (DAC) method of Yang[16], Yang and Lee,[17] Dixon and Merz,[18] and Gogonea *et al.*,[73] the adjustable density matrix assembler (ADMA) approach method of Exner and Mezey,[69] the molecular fractionation with conjugate caps (MFCC) approach developed by Zhang and co-workers,[49] and the fragment molecular orbital (FMO) method of Kitaura and co-workers.[7,46,47] Most applications of these methods to protein systems has been mostly limited to semiempirical, HF and DFT calculations. Among these approaches, FMO has been applied to higher *ab initio* level calculations such as second-order Møller-Plesset perturbation theory (MP2)[48] and coupled cluster theory (CC).[80] Moreover, the Polarizable Continuum Model[143] (PCM) has been combined with FMO approach to incorporate solvation effects in an efficient way.[144]

The FMO2-MP2 method (in conjunction with PCM) is based on a two-body expansion, which makes it substantially faster than full system calculations. Furthermore, the fragment based FMO-MP2/PCM approach has reduced memory and disk requirements which makes all-electron *ab initio* quantum mechanical calculation on macromolecules possible.[7] In order to validate the FMO scheme, a recent FMO-MP2/6-31(+)G* study based on two-body expansions showed that the error in the correlation energy relative to standard MP2/6-31(+)G* calculations was only 2.1kcal/mol error for Trp-cage.[48] Therefore, we have chosen the FMO-MP2/PCM method for our present calculations. Our goal is to validate that *ab initio* HF and MP2 methods can discriminate between native protein structures relative to a set of decoy structures. Simultaneously, we investigated how the electron correlation energy and dispersion energy varies between the native state of a protein and its corresponding decoy set. Our study is the first large-scale application of correlated *ab initio* methods to the study of protein decoy detection.

5.2 Computational Approach

5.2.1 Ab Initio Calculation

Our goal is to find an energy "scoring" function for proteins that can discriminate the native protein structures from their decoys. More specifically, the total energy of a native structure should be lower than all decoys,[26] and an energy gap, which well separates the native state(s) from the misfolded states, should be observed. Effective free energy functions have been reported in previous decoy studies;[25-30] however, the evaluation of physics-based potentials was limited to molecular mechanics (MM) and semiempirical methods. Herein, we present a correlated *ab initio* study of decoy detection. The energy-based scoring function we use to evaluate the relative stability of the protein structures is:

$$\Delta G_{tot} = \Delta E_{\mathrm{int}ra} + \Delta G_{solv} \qquad (5\text{-}3)$$

where $\Delta E_{\mathrm{int}ra}$ and ΔG_{solv} represent the intra-molecular energy (the sum of the electronic and nuclear-nuclear repulsion energies) and the solvation energy of the protein, respectively. The fragment molecular orbital method (FMO) was used to calculate the total energy of the protein $\Delta E_{\mathrm{int}ra}$ at the HF and MP2 levels.

The solvation energy term, ΔG_{solv}, in Equation 5-3 is calculated using C-PCM[145,146] combined with the FMO2 approach (*i.e.*, FMO2/CPCM).[144] Following the same spirit of the fragmentation algorithm, the induced apparent surface charges (ASC) are predetermined self-consistently based on the one-body expansions of the electrostatic potential, followed by a single ASC calculation using the two-body expansion of the electrostatic potential to further refine the ASCs. Then the HF-FMO2 energy (Equation 2-29) and MP2-FMO2 correlation energy (Equation 2-30) are calculated in the electrostatic field of the fixed ASCs.

An optimized subsystem partition scheme with a suitable buffer size for real three dimensional protein systems in DC-MP2 approach still needs to be validated. Therefore, in current study on protein decoy detection, we have chosen available FMO program implemented in GAMESS-US[38] to calculate the protein energy on HF/6-31G* and MP2/6-31G* levels. An efficient DC-HF and DC-MP2 program with highly parallel efficiency is our long-term research goal. Here we utilized FMO2-MP2 energy as the scoring function to provide some preliminary results for protein structure prediction. The C-PCM calculations used 240 tesserae per sphere and the following atomic radii:[147] $R_H = 0.01$ Å, $R_C = 1.77$ Å, $R_N = 1.68$ Å, $R_O = 1.59$ Å, $R_S = 2.10$ Å. All the solvation energies included cavitation energy contributions and van der Waals interactions between the solvent and solute.

The 6-31G* basis set was chosen for our calculations. Geometry optimization based on MP2/6-31G* gives reasonable molecular structures as shown in previous studies.[139] It is known that MP2 is able to describe the dispersion energy, but the quality of the results depends on the basis set used as well. MP2 with large basis sets overestimates the correlation interaction energy for some clusters.[148] Nevertheless, for other clusters the MP2 correlation interaction energy is close to the best estimated value given by CCSD(T) theory.[137] MP2/6-31G* usually underestimates the correlation interaction energy[142] and suffers from basis set superposition error (BSSE) due to the incompleteness. When we use relatively small basis sets such as 6-31G* in biological study here, there is no affordable way to eliminate the BSSE. MP2/6-31G* without BSSE correction always lowers the interaction energy compared to the "real" physical interaction values given by MP2/6-31G*. To illustrate these features, we have investigated two small molecule complexes: the methane dimer and the methane-benzene complex. *Ab initio* calculations using various basis sets were carried out to compute the interaction energies for these two complexes using the Qchem program.[72]

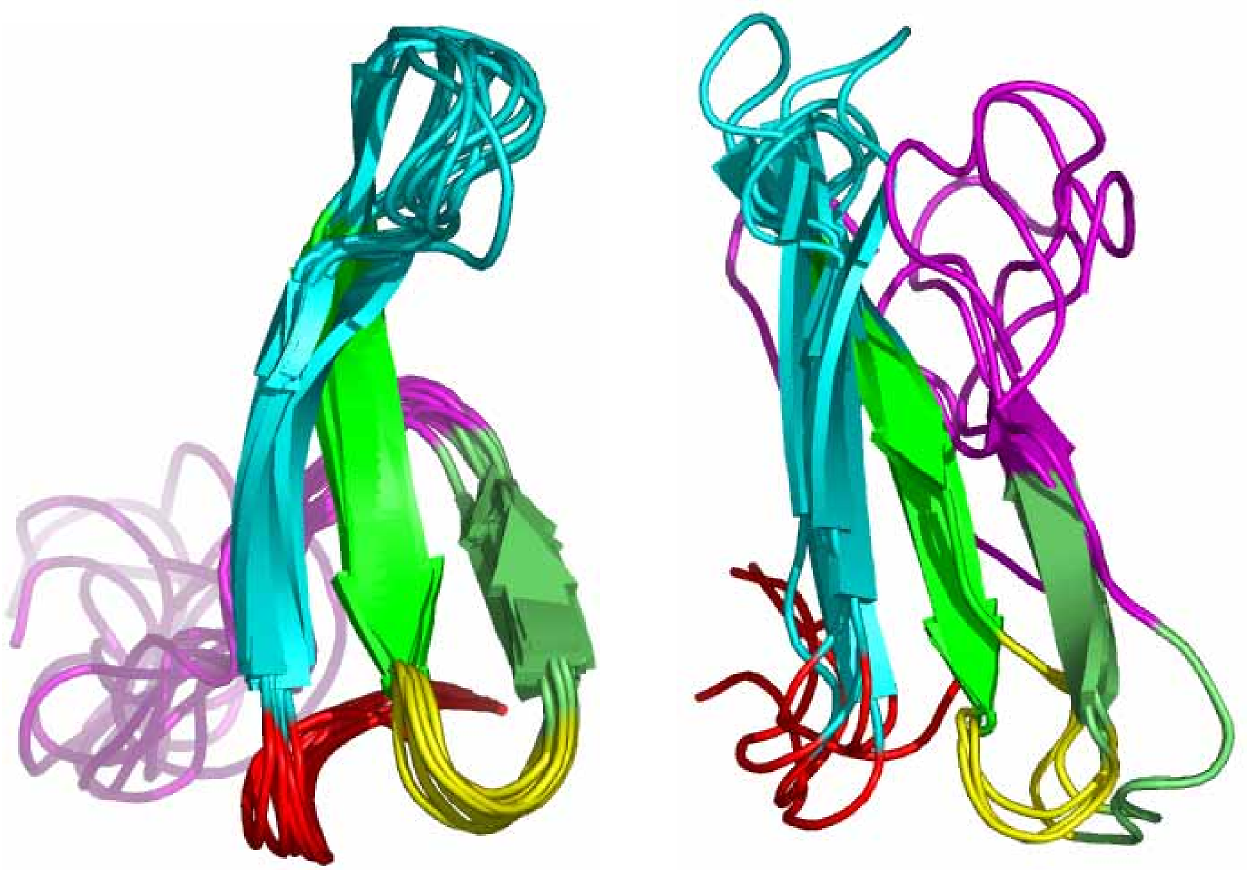


Figure 5-1. The NMR structures of the Pin1 WW domain are shown on the left, while five representative decoy structures generated by Rosetta are given on the right side of the figure. Each color denotes the same fragment for different conformations (red: residues 1-5; cyan: residues 6-16; green: residues 17-21, yellow: residues 22-24; lime: residues 25-28; magenta: residues 29-39).

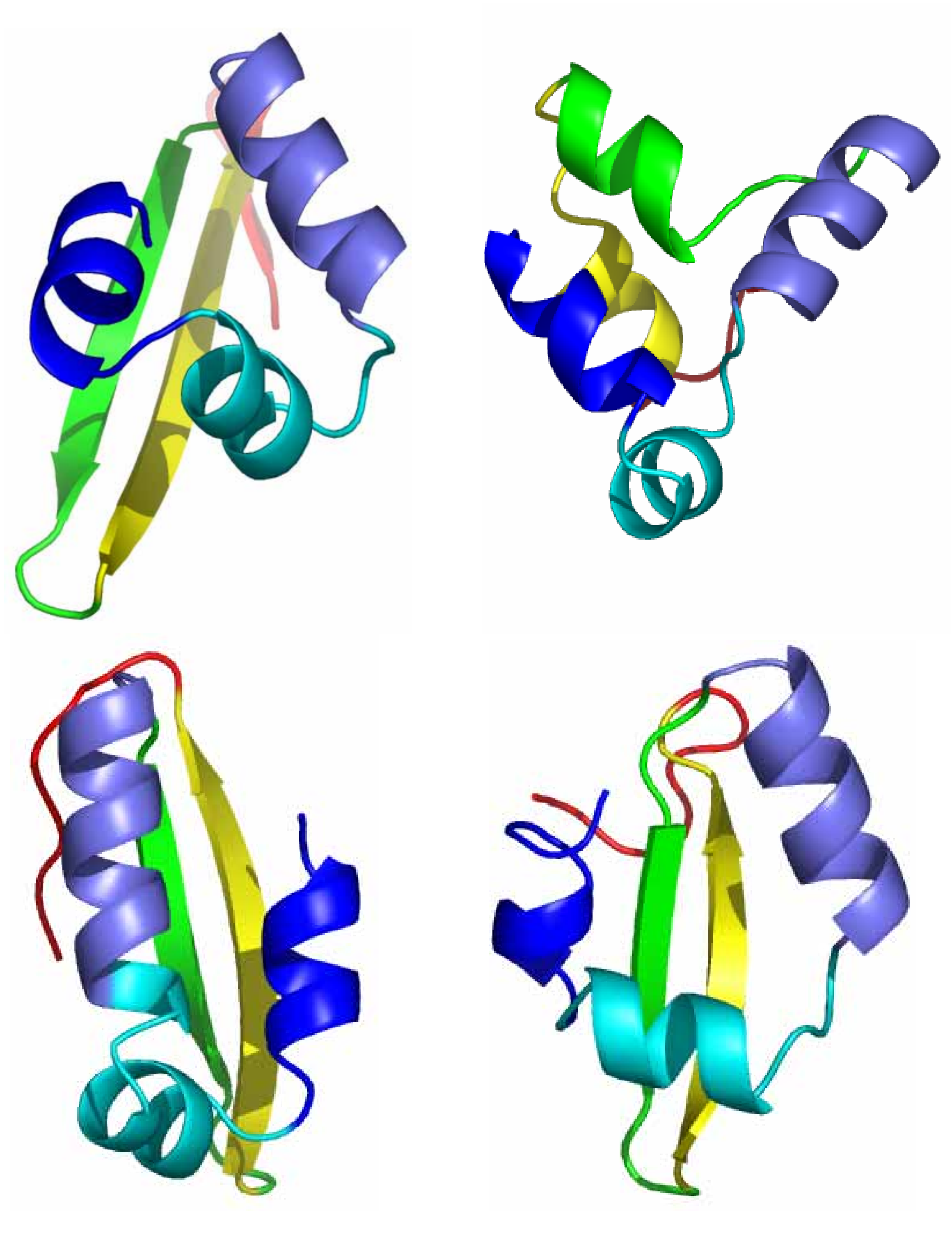


Figure 5-2. The X-ray structure of the Cro repressor (1orc, residues 7 through 57) is shown in the top left corner, while the rest three are representative decoy structures. Each color represents the same fragment for different protein conformers.

5.2.2 Decoy Selection

Nine (9) NMR structures (pdb id:1i6c) of the Pin1 WW domain were taken as our native conformations. Pin1 has 39 amino acids and contains 612 atoms in total (including hydrogen). A set of 1,000 decoy structures was generated by Rosetta.[149] Due to the relatively high expense of FMO-MP2/PCM calculations, we perform fixed radius clustering of the entire decoy set based on the mutual RMSD of C_α and C_β atoms for residues 6 through 29 using MMTSB.[150] We focused on residues 6 through 29 because this region forms an antiparallel $\beta-$sheet in the native structures while the remaining residues are in flexible loop regions (see Figure 5-1). Note that the energies we report are still for residues 1 through 39. The structures are overlaid using a least square fit before calculating RMSD values for every protein structure pair. By setting the clustering radius to 3Å, 27 subclusters were obtained. 110 structures were chosen at random from these 27 subclusters. The second protein we examined was the Cro repressor protein. The X-ray structure (pdb id: 1orc, residues 7 through 57) was taken as the native conformation and after protonation, it contained 877 atoms in total (including hydrogen atoms). Out of its Rosetta decoy set produced earlier by Baker and co-workers[151], we chose 50 decoys for this study (see below for details). Figure 5-2 shows the X-ray structure of the Cro repressor along with its three representative decoy conformations. Since it is computationally expensive to minimize all the structures at a quantum mechanical level, we performed optimizations on all of the native and decoy structures using the Generalized-Born solvation model with the AMBER FFPM3 force field [152] in order to remove bad contacts prior to the *ab initio* calculations.

5.3 Results and Discussion

5.3.1 Small Molecule Complexes

Before carrying out MP2 calculations on larger protein systems, we first investigated two small complexes in order to better understand the impact of that basis set and correlation method choices have on our computed results. The first system studied was the methane dimer. MP2 calculations were performed to derive the potential energy curves for the methane dimer using both 6-31G* and Dunning's augmented correlation consistent basis sets.[153] We also used the counterpoise correction (CP) method[154] to account for the basis set superposition error (BSSE). The energy curves with counterpoise and without counterpoise correction are shown in Figure 5-3. For a small basis set such as 6-31G*, even MP2 is unable to capture the dispersion interaction between the methane dimer after counterpoise correction. The attractive energy predicted by MP2/6-31G* without the counterpoise correction actually does not represent the physical interaction. Ironically, most of the attractive interaction energy is from BSSE emphasizing the difficulty of computing these quantities. When the basis set size is increased to Dunning's augmented correlation consistent basis sets, the dispersion energy begins to be captured at the MP2 level. In comparison to the MP2 CBS (Complete basis set method) energy at the equilibrium geometry, MP2/aug-cc-pVDZ without the counterpoise correction overestimates the dispersion energy by 0.43 Kcal/mol (88% of the MP2 CBS energy) and has a large BSSE of 0.53 Kcal/mol (108%). As the basis set increases to aug-cc-pVTZ and aug-cc-pVQZ, the potential energy curve with the CP correction converges to the MP2 CBS energy, and, not unexpectedly, BSSE decreases as the basis set increases.

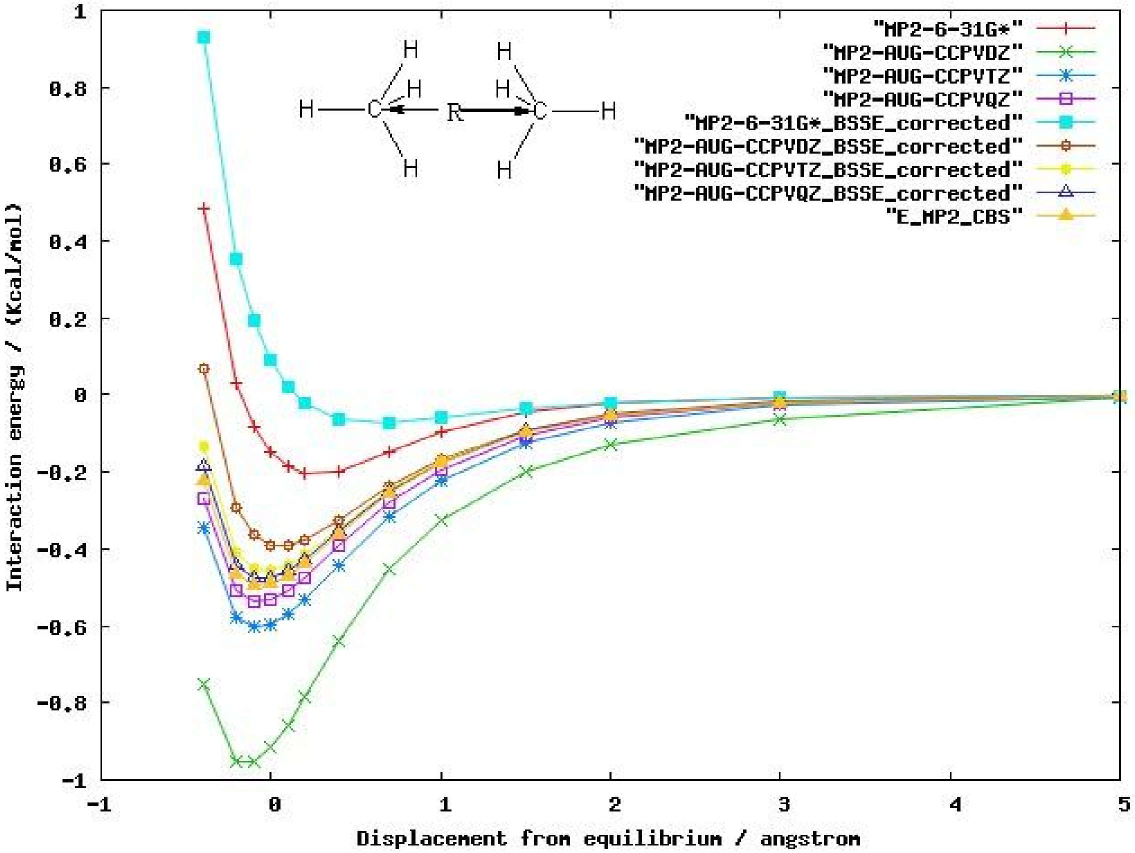


Figure 5-3. MP2 interaction energy curves for the methane dimer as a function of the center of masses (COM) distance between each methane molecule using various basis sets.

We further compare the potential energy curves by using different *ab initio* methods and the generalized AMBER force field (GAFF) [117] in Figure 5-4. Because the counterpoise correction method cannot be applied to our protein calculations, we compare the energy curves without BSSE correction using HF/6-31G*, B3LYP/6-31G* and MP2/6-31G*. The HF and DFT/B3LYP calculations, as expected, fail to capture the dispersive interaction between the methane dimer. At the equilibrium configuration, the interaction energy given by MP2/6-31G* without BSSE correction is -0.15 Kcal/mol, which underestimates the dispersion energy by 0.38 Kcal/mol (82%), when compared to

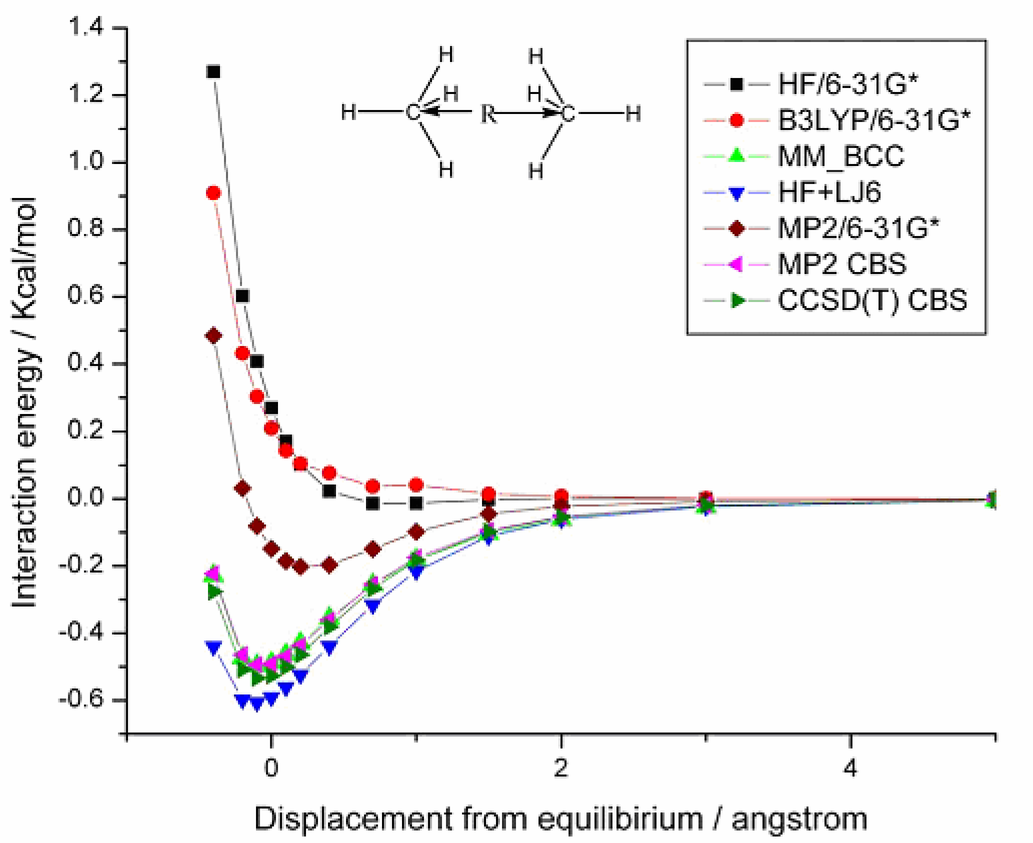


Figure 5-4. Comparison of interaction energy curves for the methane dimer as a function of the COM distance between each methane molecule at different levels of theory. See text for further details.

the -0.53 kcal/mol value obtained by CCSD(T) CBS at the equilibrium geometry. The energy curves obtained by MP2 CBS and CCSD(T) CBS are very similar to each other. The AMBER force field potential energy curve is in good agreement with the CCSD(T) CBS results, indicating that the van der Waals parameters of this complex are finely tuned.[155,156] Moreover, by adding the attractive term of the Lennard-Jones energy to the HF energy curve labeled as HF+LJ6 in Equation 5-4, the energy curve reproduces the CCSD(T) CBS energy curve, which demonstrates that the attractive term of AMBER force field compensates for the dispersion energy that is missing in the Hartree-Fock energy.

$$\Delta E_{HF+LJ6} = \Delta E_{HF} + \sum LJ6 \tag{5-4}$$

Dispersion is a pure electron correlation effect originating from the weak attractive interaction between an instantaneous dipole moment on one site and induced dipole moment on another site of the system.[157] The dipole-induced-dipole interaction is proportional to $\frac{1}{R^6}$ for large intermolecular separations.[158] Recent studies have developed dispersion corrected semiempirical[159], HF[157,160,161] and DFT[162-164] methods to remedy this problem in a pragmatic way. The dispersion corrected total energy is

$$E_{total} = E_{SCF} + E_{disp} \tag{5-5}$$

where E_{SCF} is the semiempirical, HF or DFT total energy using traditional self-consistent-filed(SCF) procedure. E_{disp} is an empirical dispersion potential given by:

$$E_{disp} = -S_6 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{C^{ij}}{R_{ij}^6} f_{damp}(R_{ij}) \tag{5-6}$$

Here, N is the number of atoms in the system, R_{ij} and C^{ij} denote the distance and dispersion coefficient between atom pair ij, respectively. $f_{damp}(R_{ij})$ is a damping function used to avoid singularities when the distance $R_{ij} \to 0$. S_6 is a global scaling factor. Thus, the dispersion energy can be evaluated in negligible computational time, which is an advantage over, more computationally expensive, non-empirical electron correlation methods. Nevertheless, same as for all other empirical methods, in order to obtain universal dispersion coefficients for different atom pairs, a thorough validation on numerous systems needs to be carried out.[142] In this study, we use the Lennard-Jones parameters from the AMBER force field.

$$E_{disp} = \sum LJ6 = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{C^{ij}}{R_{ij}^{6}} \tag{5-7}$$

Note that a damping function is not used here and, in addition, we did not scale the dispersive energy by any global scaling factor. In the AMBER LJ paramerization procedure, both the charge model RESP[165] (restrained electrostatic potential) at HF/6-31G* and AM1-BCC[166] (bond charge correction) are designed to match the electrostatic potential obtained at the HF/6-31G* level.[156] As a result, the Lennard-Jones parameters are suitable to be used with HF/6-31G* calculations.

Shibasaki *et al.* have experimentally and theoretically determined the interaction energy between methane and benzene.[167] In their calculations, the BSSE was corrected in all calculations using the CP method. We compare our computed interaction energies via MP2 calculation with CP and without CP correction in Figure 5-5. It shows features similar to those observed for the methane dimer. MP2/6-31G* with counterpoise correction has an attractive energy of -0.13 Kcal/mol, which is only 7.1% of the total dispersion energy of -1.82 Kcal/mol calculated using MP2 CBS. MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ without CP correction overestimate the dispersion energy by 80% and 32%, respectively. MP2/aug-cc-pVDZ with CP correction underestimates the dispersion energy by 0.35 Kcal/mol (19%) for the geometry at equilibrium. Until the basis set increases to aug-cc-pVTZ and aug-cc-pVQZ, the energy curves with CP correction are almost identical to the MP2 CBS results. Here the curve generated by MP2/6-31G* calculations without CP correction is very close to MP2/aug-cc-pVDZ with CP correction. We further compare the results with HF, DFT, MP2, GAFF (AMBER force field), MP2 CBS and CCSD(T) CBS in Figure 5-6. Again, HF and DFT/B3LYP fail to capture the dispersion energy for this complex. For the equilibrium geometry, the

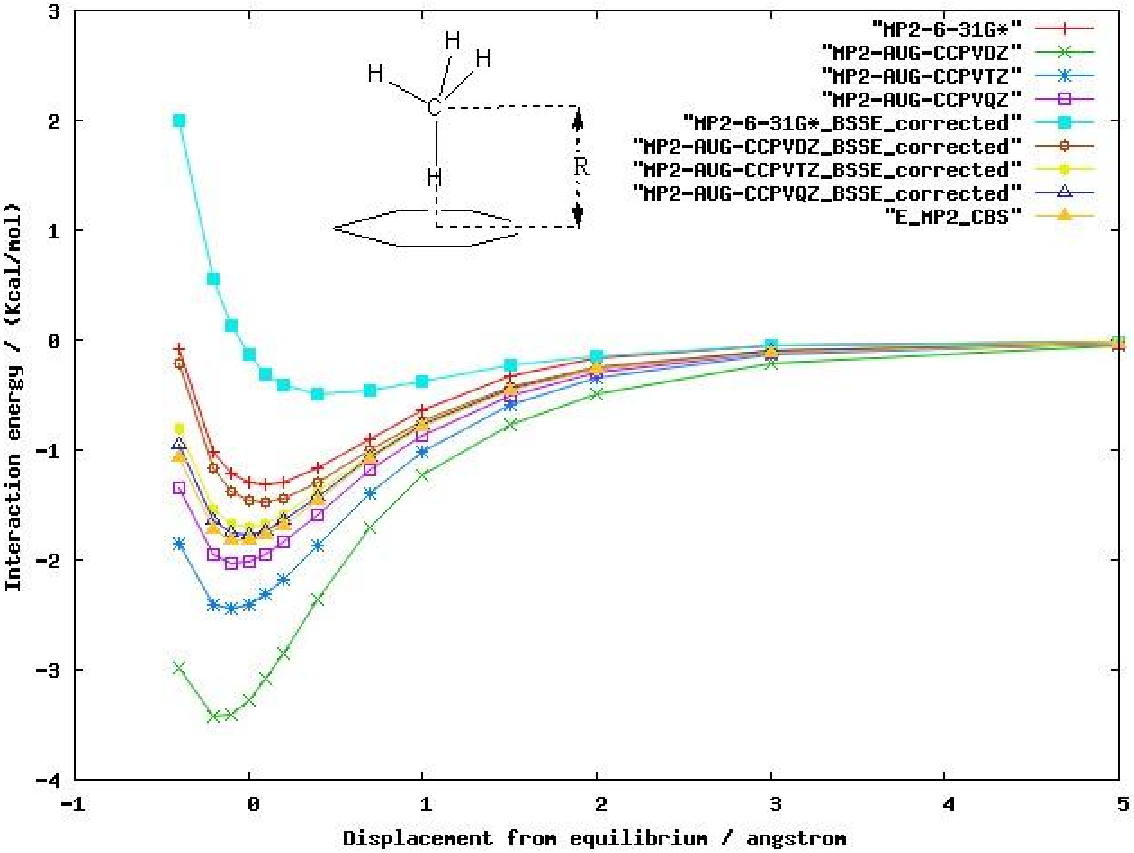


Figure 5-5. MP2 interaction energy curves for the benzene-methane dimer as a function of the COM distance between benzene and methane using various basis sets.

interaction energies are -0.92 kcal/mol, -1.30kcal/mol, -1.82kcal/mol, -1.48kcal/mol given by GAFF, MP2/6-31G* without CP correction, MP2 CBS and CCSD(T) CBS, respectively. MP2/6-31G* without CP correction fortuitously gives 88% of the total dispersion energy evaluated using CCSD(T) CBS. Most of the attractive energy originates from BSSE, rather than from dispersion. GAFF gives a qualitatively correct potential energy curve for this complex, but it underestimates the dispersive energy by 0.56 kcal/mol. In this case, MP2/6-31G* without CP correction gives a deeper energy minimum than GAFF. We also tested the performance of force fields for hydrogen bonding interactions and these results will be reported elsewhere.[168]

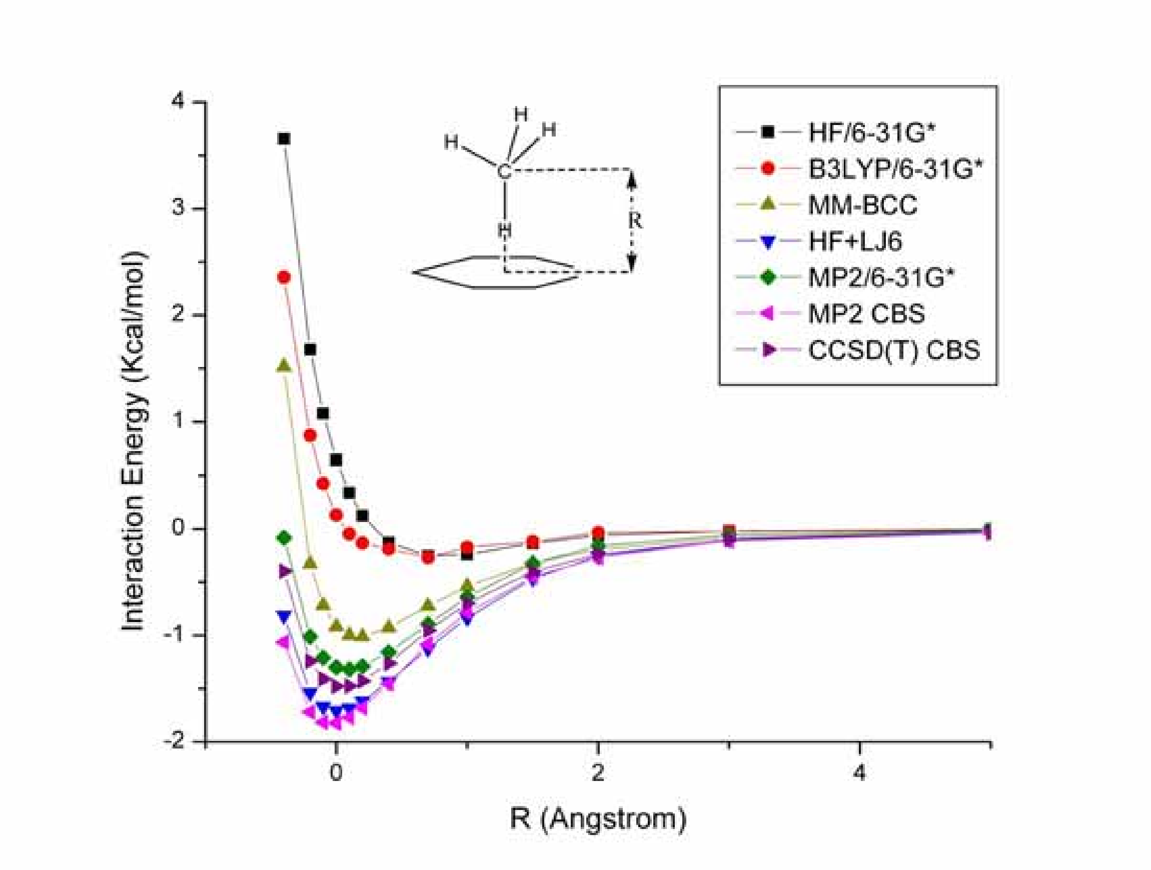


Figure 5-6. Comparison of interaction energy curves for the benzene-methane dimer as a function of the COM distance between each molecule at different levels of theory. See text for further details.

As will be shown below dispersion-dominated interactions summed over an entire protein are significant. This attractive energy has a large contribution to the total free energy of the system. Minor errors in the computed dispersion energies from large number of non-covalent interactions present in a protein will result in a deviation from the “exact” energy. The challenges faced for small molecule clusters, as summarized above, helps to set the stage for our studies using similar methods on larger macromolecules.

5.3.2 Protein Decoy Detection

Based on Equation 5-3, the HF scores of the native NMR structures (pdb id:1i6c) are higher than for most of the decoy conformations in the decoy set as shown in Figure

5-7a. The average HF energy is -584.7 kcal/mol among the native structures, while the average energy of the decoy set is -644.6 kcal/mol. Perhaps this was not too surprising because HF theory doesn't capture the dispersion energy in protein systems as illustrated in the previous HF calculations on the methane-methane and methane-benzene complexes. To compensate for this deficiency in the HF "scoring function", we added the atom-typed attractive term from the AMBER Lennard-Jones potential to the HF potential energy. We label this scoring function as HF+LJ6.

$$\Delta E_{tot} = (\Delta E_{\text{int}ra} + \Delta E_{solv})_{HF} + \sum LJ_6 \qquad \text{(5-8)}$$

Figure 5-7b shows the results using the HF+LJ6 scoring function. We find that the original trend is now reversed; all the scores of the native structures are shifted to scores lower than the average score -1120.5 Kcal/mol of the decoy set.

FMO-MP2/6-31G* calculations, in conjunction with the PCM model, were also carried out on all the structures. The energy function is then evaluated by summing the MP2 energy and the solvation energy using the PCM model.

$$\Delta E_{tot} = (\Delta E_{\text{int}ra} + \Delta E_{solv})_{MP2} \qquad \text{(5-9)}$$

Figure 5-7c shows the outcome of these calculations, which turn out to be very similar to the (HF+LJ6) energies. Indeed, the two scores are well correlated as shown in Figure 5-7d (R^2 is 0.91). After overlaying the two sets of scores with a linear square fit, the average unsigned error and root mean square deviation of the (HF+LJ6) energy from the MP2 energy are 5.09 kcal/mol and 6.79 kcal/mol, respectively. One of the "native" NMR structures was ranked third lowest in the MP2 scoring function. The decoy set created by Rosetta as shown in Figure 5-1, mostly preserved the antiparallel β-sheet-like structure of the native protein for residues 6 through 29, which likely makes this a demanding test case. Compared to X-ray structures, the NMR structures are usually more difficult to discriminate from the decoy sets.[30,31] Moreover, recent efforts to fold the WW domain have proven challenging indicating the difficulty of this example even for force field based methods with extensive sampling.[169]

None of the "native" NMR structures have the lowest energy based on the MP2 calculations. From a computational perspective, some deficiencies in our current MP2 scoring function may be the source of this observation. Firstly, the FMO method based on a two-body expansion may cause a few kcal/mol error in the total energy calculation. We did not take 3-body interactions into account in this study due to the excessive computational cost. Secondly, MP2 calculations using the 6-31G* basis set may not be sufficient to capture all of the dispersive effect. For example, we showed for the methane dimer and the methane-benzene complex that MP2/6-31G* without CP correction, underestimates the correlation energy by 0.38 kcal/mol and 0.18 kcal/mol, respectively. Note that for those intramolecular dispersion-rich interactions, the attractive energy given by MP2/6-31G* is mainly attributed to BSSE, other than the real dispersion energy.[168] Thirdly, to accurately evaluate the solvation energy of a protein is still a significant challenge for theoretical chemists and the PCM model, while effective may not ultimately be the best choice. Even for small ionic species the mean unsigned errors of various theoretical models can be more than 4.0 kcal/mol compared to experimental results.[170,171] Hence, the PCM solvation model likely contributes to the observed errors in our scoring function. A final source of concern is the quality of NMR structures in general. The variability in the stability of the 9 NMR structures examined

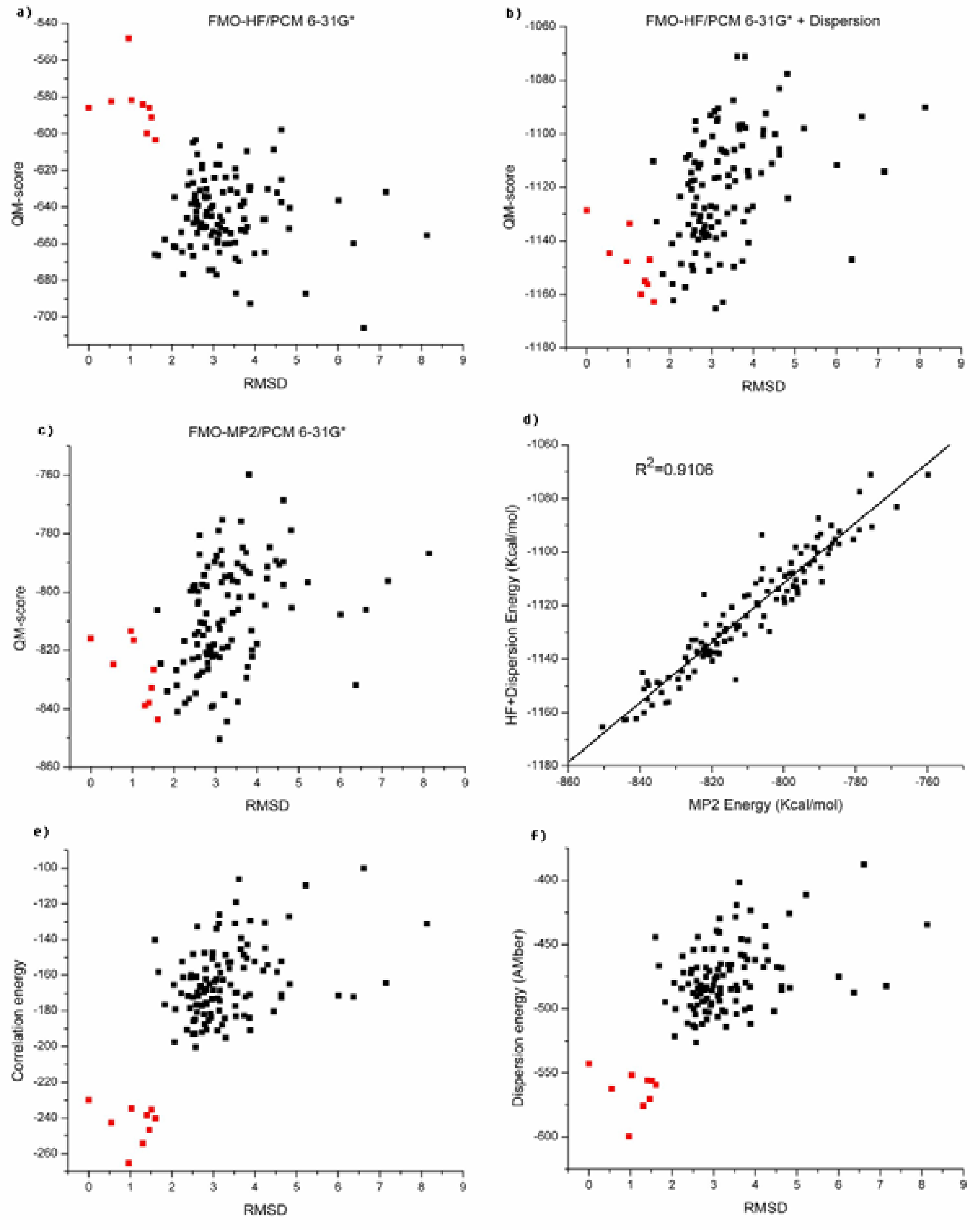


Figure 5-7. Six different energy scores for the native and decoy states of the Pin1 WW domain (1i6c). Solvation energies are included in (a),(b),(c),(d). (a) HF/6-31G*. (b) HF/6-31G*+LJ6. (c) MP2/6-31G*. (d) The correlation between MP2/6-31G* energies and (HF/6-31G*+LJ6) energies. (e) Electron correlation energies given by MP2/6-31G*. (f) The attractive term of the Lennard-Jones energies (LJ6).

here is on the order of 30 kcal/mol at the MP2/6-31G* level. We have noted issues with NMR structures in the past when using semiempirical QM scoring functions.[30]

We also extracted the electron correlation energy in the solvent by subtracting the HF energy from the MP2 energy as given in equation 5-10 to determine the role correlation plays in decoy detection.

$$\Delta E_{correlation} = (\Delta E_{\mathrm{int}\,ra} + \Delta E_{solv})_{MP2} - (\Delta E_{\mathrm{int}\,ra} + \Delta E_{solv})_{HF} = \Delta E_{MP2}^{corr}(\psi_{solv}) \qquad (5\text{-}10)$$

where ψ_{solv} denotes the ground-state wavefunction of the protein in the solvent.

We find that the electron correlation energy has significant discrimination ability between decoy and native structures (see Figure 5-7e). Likely this reflects a tighter packing of amino acids in this dispersion dominated case. This is further reinforced if we only use the dispersive term of the Lennard-Jones energy (LJ6) as a scoring function.

$$\Delta E_{dispersion} = \sum LJ6 \qquad (5\text{-}11)$$

Figure 5-7f illustrates the LJ6 scores for all the structures. The scoring function in terms of dispersion energy works as well as the electron correlation energy in this system. The energy gap between the average score of the native states and the decoys using the electron correlation scoring function is 79.3Kcal/mol, while the energy gap given by LJ6 score is 87.7Kcal/mol.

To push our analysis further we analyzed several systems in search for a case where dispersion is not the dominant driving force (as evaluated using the AMBER dispersion term). The Cro repressor (pdb id:1orc) was identified as a suitable test case for our purposes. The Rosetta decoys for this protein have already been published by Baker and co-workers[151]. In order to streamline our calculations, we first evaluated the AMBER dispersive energies of all of the 1,000 decoys and then took the 50 decoy

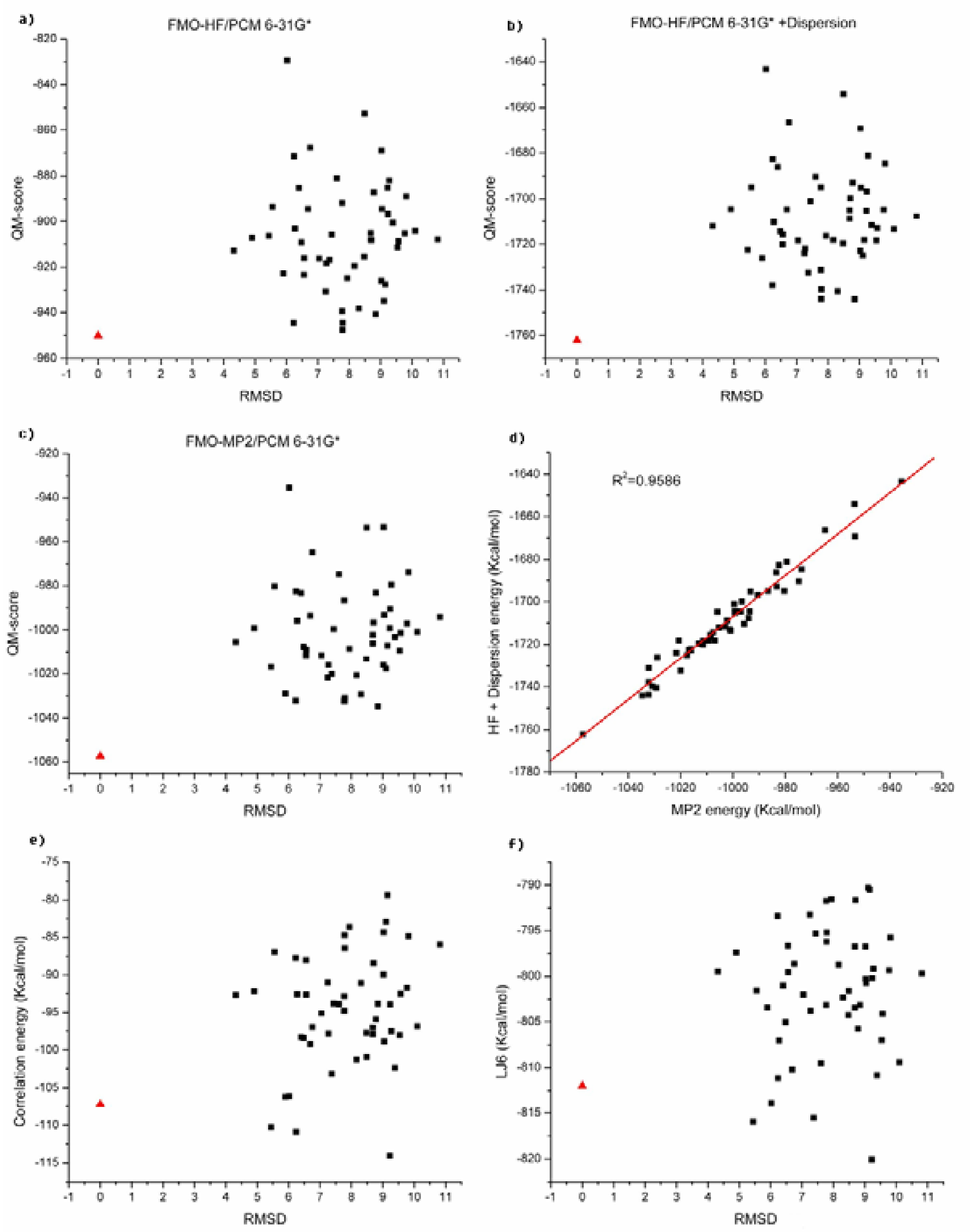


Figure 5-8. Similar to figure 5-7, but for the Cro repressor (1orc). The red triangle represents the native X-ray structure while the black squares represent the decoys.

structures which had the lowest dispersive energies when compared to the rest of the decoys. As shown in Figure 5-8f, the empirical dispersive energy $\sum LJ6$ of the native structure is ranked 5th in comparison to the decoys. The HF energy of the native structure is only 2.54kcal/mol less than the lowest HF energy of the decoy set (see Figure 5-8a). By adding the empirical dispersive term to the HF energy, the native structure becomes well separated from the decoy set by 18.1 kcal/mol compared to the lowest score of the decoys (see Figure 5-8b). The MP2 based scoring works as well as HF+LJ6 score with a difference of 22.7 kcal/mol between the native conformation and the lowest energy decoy (see Figure 5-8c). The correlation between the computed MP2 energy and HF+LJ6 energy is 0.96 (see Figure 5-8d). They are highly correlated as was observed for the Pin1 WW domain (1i6c) (see Figure 5-7d). We also extracted the *ab initio* electron correlation energy (Equation 5-10) as shown in Figure 5-8e. In this case, neither the empirical dispersive energy nor the electron correlation energy was a suitable descriptor to rank this non-dispersion dominated protein folding example correctly.

It is interesting to consider how important the choice of Lennard-Jones dispersion parameters is on ranking protein decoys. In general, these terms are finely tuned for their specific interaction types, but is this “tuning” absolutely necessary? Since these individual terms are relatively small in magnitude and do not have a radial dependence one could speculate that the choice of parameter is less important than simply providing some measure of dispersive type interactions. To further investigate this we randomly scrambled the “standard” Lennard-Jones parameters and then rescored accordingly. As shown in Figure 5-9a, after the Lennard-Jones parameters for LJ6 term were randomly scrambled for the Pin1 WW domain, the dispersive energy does not separate the 9 NMR structures from the decoy set. The energies of a few native structures are higher than some of the decoys. The correlation between the MP2 energy and the HF+LJ6 energy drops from 0.91 to 0.28 clearly indicating a degradation in the correlation(see Figure 5-9b). For the Cro repressor (1orc), the rank of the native structure drops from fifth to twelfth after the Lennard-Jones parameters were randomly scrambled (compare Figure 5-10a with Figure 5-8f). The sum of HF energy and the dispersion energy (HF+LJ6) of the native structure is only 0.88 kcal/mol less than the lowest HF+LJ6 energy decoy (see Figure 5-10b). The gap was 18.1 kcal/mol as shown in Figure 5-8b when the “correct” Lennard-Jones parameters were employed. Similar to the Pin1 WW domain (1i6c), the correlation between the MP2 energy and HF+LJ6 energy dramatically drops from 0.96 to 0.08 for Cro repressor again indicating that the MP2 energy and HF+LJ6 energy are uncorrelated when “incorrect” Lennard-Jones parameters are employed. Regardless of how the LJ6 parameters were scrambled we obtained similar results as those described here. Hence, we conclude that the nature of the LJ6 parameter set is critical to correctly detect decoys over native structures. Moreover, the correlation energy we obtain using our MP2 calculations (in so far as these terms represent effective dispersion) could be used to improve LJ6 terms employed in standard force fields through the optimization of the correlation coefficient between the MP2 and HF+LJ6 results. Overall, our study suggests that van der Waals parameters need to be carefully parameterized to experimental interaction energies or accurate *ab initio* calculations and in the case of the AMBER LJ6 set this seems to of been achieved.

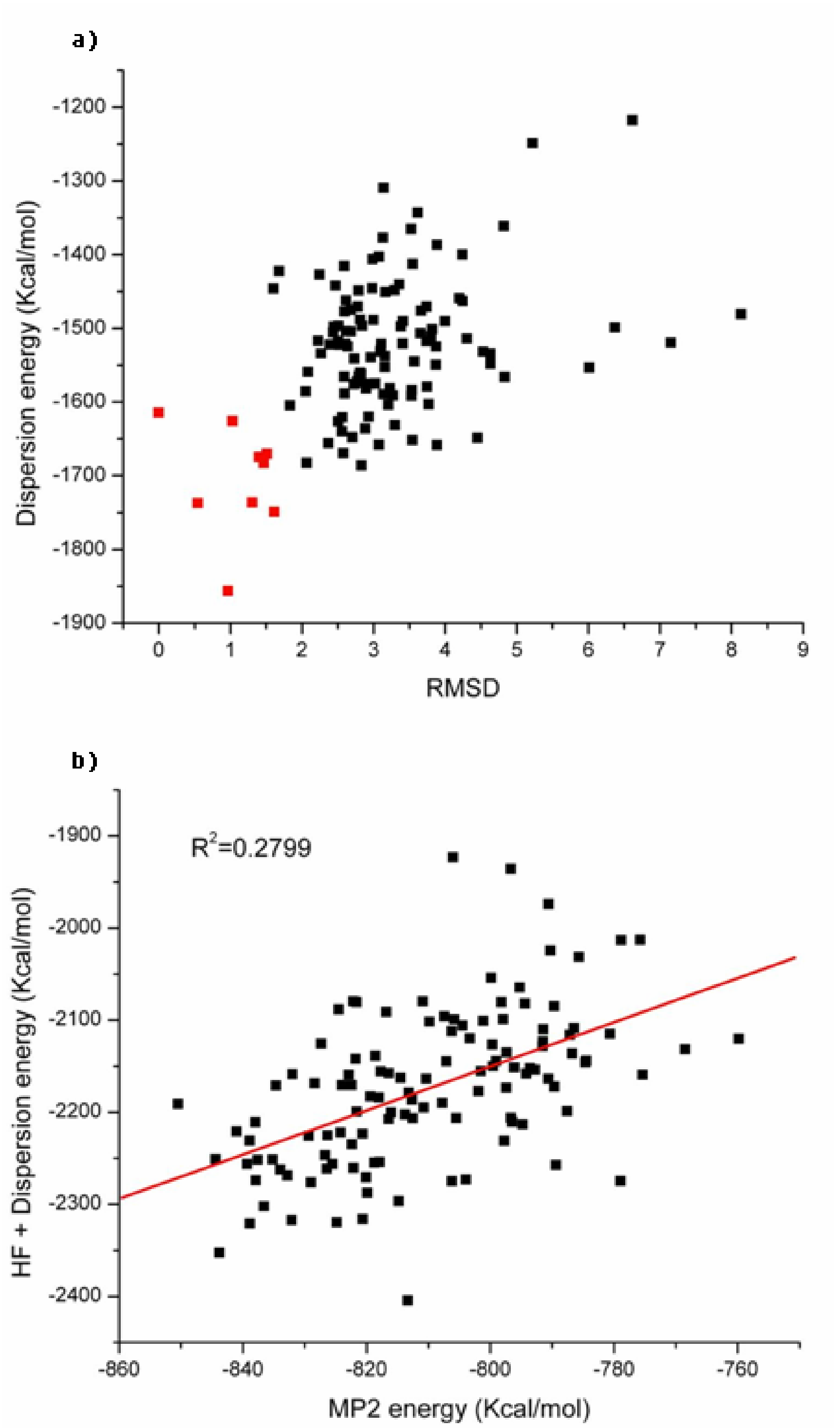


Figure 5-9. Randomly scrambled Lennard-Jones LJ6 parameters. (a) Labels are similar to Figure 5-7f for the Pin1 WW domain (1i6c). (b) The correlation between the MP2 and (HF+LJ6) energies for the Pin1 WW domain (1i6c) after the Lennard-Jones LJ6 parameters are randomly scrambled.

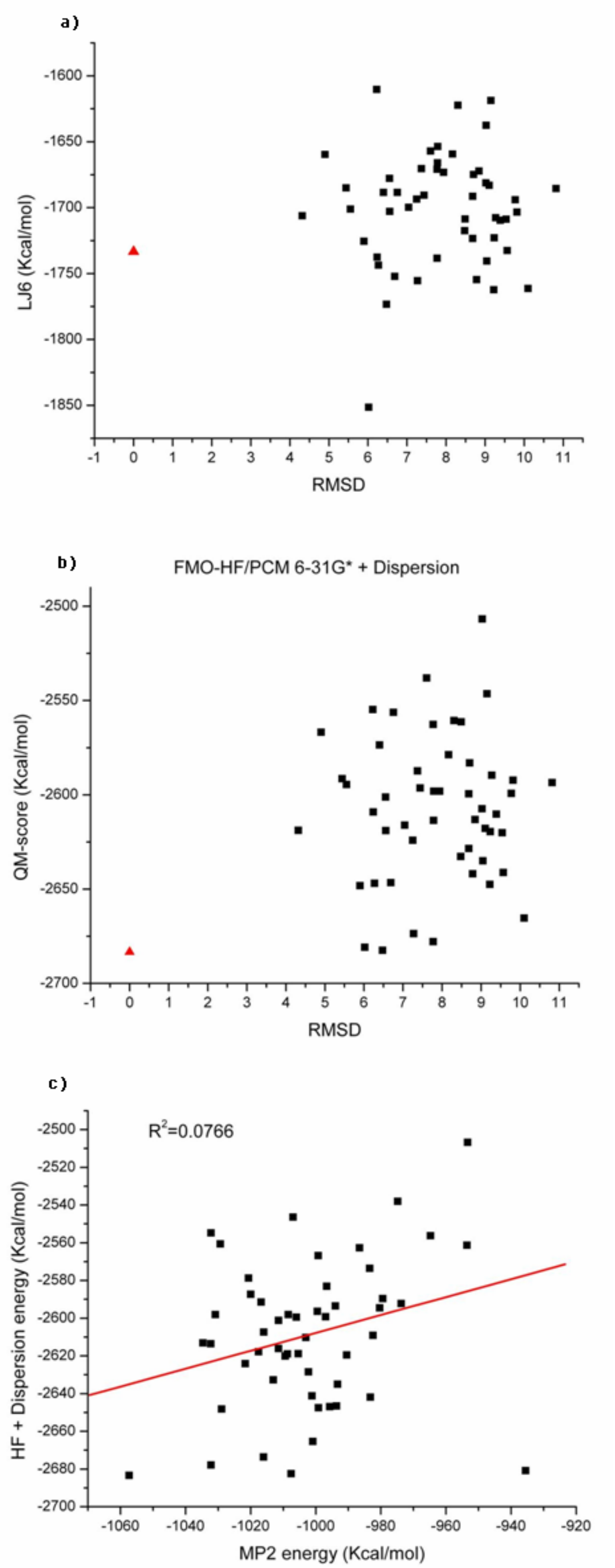


Figure 5-10. Randomly scrambled Lennard-Jones LJ6 parameters for the Cro repressor (1orc). (a) Labels are similar to Figure 5-8f. (b) Labels are similar to Figure 5-8b. (c) The correlation between the MP2 and (HF+LJ6) energies

5.4 Conclusions

In this chapter, we carried out large scale MP2 calculations on native and computer-generated decoy sets of two protein systems. The two proteins employed represent a case where dispersion appears to dominate the folding (WW domain) and one where this is less so (Cro repressor). In general, HF calculations fail to rank the native protein structures in the dispersion dominated Pin1 WW domain, because HF formally cannot capture dispersion interactions. When the MP2 correlation energy is added to the HF energy, the energies of native structures improve relative to the decoy structures. In the dispersion dominated case we studied here, the correlation energy turns out to be very good at discriminating the native NMR structures from the non-native conformations, which suggests a more favorable packing of non-polar residues in native states relative to the decoy sets. In the non-dispersion dominated Cro system, both the MP2 calculations (including solvation) as well as the HF+LJ6 calculations performed well in ranking native versus decoy structures.

Furthermore, we found that the sum of the Hartree-Fock energy and the dispersion energy given by AMBER LJ6 term correlates extremely well with our computed MP2 energies for both proteins studied. Since MP2 calculations are much more computationally intensive than HF; the HF+LJ6 energies provide a route to rapidly obtain near MP2 quality results. We also find that the nature of the Lennard-Jones parameters is critical to make this approach work. In this regard the current AMBER LJ6 parameters associated with the HF energy computed using 6-31G* basis set reproduce MP2/6-31G* trends.

The application of efficient and accurate linear-scaling *ab initio* calculations to biological systems is coming of age.[7,12,18,49,73] In the current study, single point FMO2-

HF/6-31G* PCM and FMO2-MP2/6-31G* PCM calculations on the Pin1 WW domain (1i6c) cost 12 days and 23 days on average on a single 2.4GHz AMD Opteron(tm) 250 Processor, respectively. Clearly these are still quite expensive calculations using a single processor. The FMO implementation in GAMESS is not particularly efficient and other codes (for example, QChem[72], CP2K[172], etc.) have more efficient direct SCF calculations when using a single processor. FMO is more efficient when run in parallel, but given the large number of decoys we studied we opted for the trivially parallel approach where we ran hundreds of calculations on single processors at more-or-less the same time (given the vagaries of machine crashes, power outages, etc.). Looking for more robust algorithms using linear-scaling methods clearly continues to be a very significant challenge for theoretical chemists. Furthermore, accurate solvation models are indispensable for high quality scoring functions. PCM, while quite stable and robust, underperforms other approaches available in the literature.[170,171] We also note that recently described density functionals such as PWB6K and M06-class provides good performance for interaction energies both in hydrogen-bonding and dispersion-dominated complexes.[173,174] Dispersion corrected DFT[162-164] is another alternative approach since the dispersion energy can be calculated rapidly, but the universal parameters need to be well fit using large data sets. For large calculations using the MP2 method, high quality basis sets are usually required to achieve accurate potential energies. Most of the intramolecular dispersion interaction calculated by MP2/6-31G* is attributed to the non-physical BSSE.[168] For full MP2 calculations on protein systems it is not feasible either to correct for the basis set superposition error associated with 6-31G* or to use a large basis sets such as Dunning's augmented correlation consistent basis sets. On the contrary, the (LJ6) term in dispersion corrected HF scoring function gives the real dispersion energy, therefore, (HF+LJ6) offers a more physical and affordable model to describe the potential energy for proteins.

CHAPTER 6
ACCURATE BENCHMARK CALCULATIONS ON THE GAS-PHASE BASICITIES OF SMALL MOLECULES

6.1 Introduction

For continuum based condensed-phase molecular dynamics simulations, an accurate continuum solvation model is important in order to accurately simulate the motions of atoms in the aqueous phase.[175] For many solvation models, a set of empirical parameters is finely tuned to reproduce experimental solvation free energies. In order to have a set of reliable experimental reference data, substantial effort has been devoted to compilations of solvation free energies.[170,176-180] For neutral species, Truhlar and coworkers have concluded that the uncertainty in experimental solvation free energies is typically as low as 0.2 kcal mol^{-1}.[181] On the other hand, for the aqueous solvation free energies of ionic species, a typical experimental error of 4-5 kcal mol^{-1} was estimated because of the uncertainties in associated experimental quantities.[181] Hence, the relatively large uncertainty of reference values for ionic solutes has hindered the critical assessment of current continuum solvation models.

The aqueous solvation free energies of an anion A$^-$ ($\Delta G_S^*(A^-)$) can be determined using the thermodynamic cycle shown in scheme 6-1, and $\Delta G_S^*(A^-)$ is defined as,[182]

$$\Delta G_S^*(A^-) = \Delta G_S^*(AH) + \Delta G_{aq}^*(AH) - \Delta G_S^o(H^+) - \Delta G_{bas}^o(A^-) \qquad (6\text{-}1)$$

where $\Delta G_S^*(AH)$ is the solvation free energy of the neutral species AH, $\Delta G_{aq}^*(AH)$ is equal to $2.303RTpK_a(AH)$, (where $pK_a(AH)$ is the negative common logarithm of the aqueous-phase acid dissociation constant of AH). $\Delta G_S^o(H^+)$ is the standard aqueous solvation free energy of the proton, $\Delta G_{bas}^o(A^-)$ is the gas-phase basicity of the anion A$^-$

Scheme 6-1. The thermodynamic cycle

$$\begin{tikzcd}
AH\,(g) \arrow[r, "\Delta G_g^o(AH)"] \arrow[d, "\Delta G_s^*(AH)"] & A^-\,(g) \arrow[d, "\Delta G_s^*(A^-)"'] & + & H^+\,(g) \arrow[d, "\Delta G_s^*(H^+)"] \\
AH\,(aq) \arrow[r, "\Delta G_{aq}^*(AH)"'] & A^-\,(aq) & + & H^+\,(aq)
\end{tikzcd}$$

defined as

$$\Delta G_{bas}^{o}(A^-) = G_{gas}^{o}(A^-) + G_{gas}^{o}(H^+) - G_{gas}^{o}(AH) \qquad (6\text{-}2)$$

Kelly *et al.* have reported the estimated uncertainties for the solvation free energy of anions ($\Delta G_S^*(A^-)$) using the root-sum-of-squares combinations of the experimentally measured quantities on the right side of the equation 6-1.[170] The typical uncertainty of the solvation free energy of anions is 2-3 kcal mol^{-1}. An average uncertainty of 0.2 kcal mol^{-1} for the solvation energy of neutral solutes ($\Delta G_S^*(AH)$) was previously estimated.[181] The experimental $pK_a(AH)$ within the range of 0-14 can be measured fairly precisely, therefore, the uncertainty of $\Delta G_{aq}^*(AH)$ is negligible for the estimation the overall uncertainty of $\Delta G_S^*(A^-)$. For the aqueous solvation free energy of the proton, Kelly *et al.* assigned an uncertainty of 2 kcal mol^{-1},[170] which has a large contribution to the overall uncertainty of $\Delta G_S^*(A^-)$. The gas-phase basicities of the anions $\Delta G_{bas}^{o}(A^-)$ were originally taken from the NIST standard reference database[183]. In this study, we took the values and their uncertainties from the data sets collected by Kelly *et al.*[170]. For several anions, there is more than one experimental measurement available, and a typical uncertainty of 2 kcal mol^{-1} is assigned for most of the anions.[184-187] For some cases, the uncertainties of the gas-phase basicities are as large as 2.8 kcal mol^{-1}, which significantly increases the overall uncertainties of the solvation free energies of anions.

During the past two decades, great progress has been made towards achieving the goal of predicting thermodynamic properties to "chemical" accuracy (1 kcal mol^{-1}).[188,189] High-level electron correlation theory, *e.g.* CCSD(T)[190] incorporating high angular momentum basis functions has become the "gold standard" approach for obtaining thermochemical properties to chemical accuracy. Higher accuracy can be further attained by extrapolation of the energies to the complete basis set limit (CBS).[191,192]

Previous studies[193-222] have been carried out to calculate the gas-phase basicities and acidities of molecules. Burk and co-workers,[199,201] Koppel *et al.*[194] have critically assessed the performance of density functional theory for prediction of gas-phase acidities and basicities. Burk *et al.* have concluded that the average absolute errors can fall below 2.5 kcal mol^{-1} for their test sets (49 acids and 32 bases) based on B3LYP/6-311+G(3df,3pd) calculations.[199] Many-body perturbation theory (MBPT)[223] and coupled-cluster theory (CC)[224-228] in conjunction with G2[229], G3[230] and "multi-level" approaches (e.g. CBS-QB3[210,231], G3B3[232], G3MP2B3[232], MCCM/3[233] and SAC/3[233]) have been proposed to obtain thermochemical data to chemical accuracy. In these procedures, a series of calculations are carried out at different levels of theory with different basis sets. Zero-point energy and high-level corrections were made based on the additivity approximation. For instance, the CBS-QB3 theory optimizes the geometries of molecules and calculates thermochemical data at the B3LYP/6-311G(2d,d,p) level, followed by a series of MP2, MP4 and CCSD(T) calculations using Pople type basis sets to obtain the electron correlation energy. Ervin and Deturi have found that CCSD(T)/aug-cc-pVTZ calculations give more accurate gas-phase acidities than CBS-QB3 theory for the molecules they tested,[193] which indicates that large basis sets are required to obtain accurate electron correlation energies of molecules. However, CCSD(T) calculations using aug-cc-pVTZ are limited to small molecules due to the poor scaling properties (N^7, where N is the number of basis functions) for CCSD(T) calculations. In addition, they did not extrapolate the CCSD(T) energies to the complete basis set limit.

Martin and co-workers have developed the W1 and W2 methods,[205,207] where the CCSD and CCSD(T) energies are extrapolated to the infinite-basis limit. Moreover, contributions from inner-shell correlation, scalar relativity, atomic spin-orbit splitting and anharmonic zero-point energies were also included. One of the most sophisticated computations which have been done so far is by Allen and co-workers.[200] They have performed all-electron coupled-cluster (AE-CC) calculations up to single, double, triple, quadruple and pentuple excitations with Dunning's augmented correlation-consistent, atom-centered Gaussian basis sets. They have also included the core electron correlation, scalar relativistic effects, diagonal Born-Oppenheimer corrections (DBOC)[234-237] and anharmonic zero-point energies. However, such expensive calculations are currently limited to molecules with 2 heavy atoms and serve more as benchmark calculations rather than as an approach that can be applied generally.

It is well known that accurate calculation of the electron correlation energy requires a large atom-centered Gaussian basis set. In this chapter, we use Dunning's augmented correlation-consistent basis sets (aug-cc-pV*n*Z)[153,238,239] (where *n*=D,T,Q) for benchmark MP2 and CCSD(T) calculations on gas-phase basicities and extrapolate the results to the complete basis set limit. Thereby, the errors arising from the incompleteness of the basis can be largely reduced.[240] The goals of this study are (1) to benchmark the accuracy of different *ab initio* theories (HF, MP2 and CCSD(T)) for the theoretical estimation of the gas-phase basicities of molecules and (2) to identify an efficient approach which is able to achieve chemical accuracy for gas-phase basicity calculations on systems containing up to 10 heavy atoms. We can use the resultant approach as a useful computational protocol to validate experimental gas-phase basicities, when more than one experimental measurement is available, and to even make accurate theoretical estimates for the cases where experimental values are not available. In this study, we include some unusual molecules, such as hydroperoxides, in the test set of 41 molecules; furthermore, we have also examined the conformational effects for accurately theoretical prediction of gas-phase basicities.

6.2 Computational Details

We used the Gaussain03 package[110] for all *ab initio* calculations. MP2/aug-cc-pVTZ calculations were carried out on all the molecules for geometry optimizations, vibrational frequencies and thermochemical analyses. The zero-point vibrational energies (ZPVEs) only include harmonic contributions. Subsequently, frozen-core MP2 and CCSD(T) single point energy calculations using augmented correlation-consistent basis sets (aug-cc-pV*n*Z) were employed on the optimized structures. The two point extrapolation scheme[191]

$$E_{MP2\_CBS} = E_{MP2,x} + \text{constant} \times x^{-3} \qquad (6\text{-}3)$$

was used to obtain the complete basis set (CBS) extrapolated values of the MP2 correlation energies (E_{MP2_CBS}) from energy calculations using two different basis sets, aug-cc-pVTZ and aug-cc-pVQZ, . The variable **x** in Equation 6-3 represents their largest angular momentum of the basis set, i.e. x=3 for aug-cc-pVTZ and x=4 for aug-cc-pVQZ. The Hartree-Fock energies were not extrapolated and were simply taken from the results of the larger basis set (aug-cc-pVQZ) calculations. The CBS correlation energies for CCSD(T) were obtained using:

$$E_{CCSD(T)\_CBS} = E_{MP2\_CBS} + (E_{CCSD(T),aug\text{-}cc\text{-}pVDZ} - E_{MP2,aug\text{-}cc\text{-}pVDZ}) \qquad (6\text{-}4)$$

which is based on the observation that the difference between the MP2 and CCSD(T) correlation energies converges faster in basis set size than the correlation energies themselves[241-243]. The effectiveness of the computational approach shown in Equation 6-4 is based on the propositions within the so-called focal-point analysis (FPA) scheme**.**[200,244-246] The internal thermal energy corrections (translational E_{trans}, rotational E_{rot} and vibrational E_{vib}) were made to the electronic energy,[247]

$$E_{tot} = E_{elec} + E_{trans} + E_{rot} + E_{vib} \qquad (6\text{-}5)$$

The Gibbs free energy G was calculated from

$$H=E_{tot}+RT \qquad (6\text{-}6)$$

$$G=H-TS_{tot} \qquad (6\text{-}7)$$

Where R is the gas constant, T is the temperature, H is the enthalpy and $S_{tot} = S_{trans} + S_{rot} + S_{vib} + S_{elec}$ (contributions from translational, rotational, vibrational and electronic motions, respectively). The gas-phase basicity of a species A^- is defined in Equation 6-2. The standard state was 298.15 K and 1 atm pressure.

6.3 Results and Discussion

6.3.1 Gas-phase Basicity Calculations

First, to assess the accuracy of the complete basis set limit for MP2 and CCSD(T) calculations, we carried out full *ab initio* CCSD(T)/aug-cc-pVTZ and CCSD(T)/aug-cc-pVQZ calculations on five small molecules (H_2O, H_2S, HCN, C_2H_2, H_2O_2) for comparison. One can see from Table 6-1, for the same optimized geometries obtained from MP2/aug-cc-pVTZ calculations, HF/aug-cc-pVQZ has the largest RMSE of 5.6 kcal mol^{-1} compared to experimental values. MP2/aug-cc-pVQZ, MP2_CBS (MP2 with complete basis set estimate) and CCSD(T)/aug-cc-pVDZ results have smaller RMSEs between 2.0 kcal mol^{-1} and 2.6 kcal mol^{-1}. CCSD(T)_CBS (CCSD(T) with complete basis set estimate) performs just as well as the significantly more expensive CCSD(T)/aug-cc-pVTZ and CCSD(T)/aug-cc-pVQZ levels. Note that the CCSD(T)_CBS results are extrapolated from MP2_CBS and CCSD(T)/aug-cc-pVDZ calculations with no additional computational cost. Due to the poor scaling of CCSD(T), it is not economical to calculate the Gibbs free energy for relatively larger molecules using large basis sets such as aug-cc-pVTZ and aug-cc-pVQZ, however, the extrapolation using Equation 6-4 strikes a compromise between the computational expense incurred and the attained accuracy for our test on five representative small molecules.

Next, we applied the extrapolation approach using Equation 6-4 for the remaining 36 molecules and the results are shown in Table 6-2. HF/aug-cc-pVQZ has the largest overall RMSE for this test set. MP2/aug-cc-pVQZ and MP2_CBS have similar performance with very close RMSEs of 3.0 kcal mol^{-1} and 3.2 kcal mol^{-1}, respectively. CCSD(T)/aug-cc-pVDZ outperforms the MP2 results, with a RMSE of 2.2 kcal mol^{-1}. Among all the approaches we tested, CCSD(T)_CBS has the lowest RMSE of 1.0

Table 6-1. Calculated and experimental gas-phase basicities of five representative small molecules (in kcal mol^{-1}).*

	HF/	MP2/	CCSD(T)/	MP2	CCSD(T)	CCSD(T)	CCSD(T)	Exp.[183]
	aug-cc-	aug-cc-	aug-cc-	_CBS	_CBS	/aug-cc-	/aug-cc-	
	pVQZ	pVQZ	pVDZ			pVTZ	pVQZ	
H_2O	393.7	380.0	381.9	379.8	383.7	384.1	384.3	383.7±0.2
	(+10.0)	(-3.7)	(-1.8)	(-3.9)	(0.0)	(+0.4)	(+0.6)	
H_2S	346.8	342.9	343.8	342.4	345.5	345.5	345.2	344.9±1.2
	(+1.9)	(-2.0)	(-1.1)	(-2.5)	(+0.6)	(+0.6)	(+0.3)	
HCN	342.5	342.4	340.8	342.3	343.1	343.4	343.3	343.7±0.3
	(-1.2)	(-1.3)	(-2.9)	(-1.4)	(-0.6)	(-0.3)	(-0.4)	
H_2O_2	375.4	368.0	367.6	367.8	369.2	368.9	369.1	368.6±0.6
	(+6.8)	(-0.6)	(-1.0)	(-0.8)	(+0.6)	(+0.3)	(+0.5)	
C_2H_2	372.7	368.9	365.5	369.0	369.5	369.4	-	370.0±1.8
	(+2.7)	(-1.1)	(-4.5)	(-1.0)	(-0.5)	(-0.6)		
MAXE	10.0	-3.7	-4.5	-3.9	0.6	0.6	0.6	
MSE	4.0	-1.7	-2.3	-1.9	0.0	0.1	0.3	
MUE	4.5	1.7	2.3	1.9	0.5	0.4	0.5	
RMSE	5.6	2.0	2.6	2.2	0.5	0.5	0.5	

*For the five columns (HF/aug-cc-pVQZ, MP2/aug-cc-pVQZ, CCSD(T)/aug-cc-pVDZ, MP2_CBS, CCSD(T)_CBS), geometry optimizations and thermochemical analyses were all performed at MP2/aug-cc-pVTZ level. The ZPVEs only include the harmonic contributions. The electronic energies on the optimized geometries were calculated at HF/aug-cc-pVQZ, MP2/aug-cc-pVQZ, CCSD(T)/aug-cc-pVDZ, and extrapolated to complete basis set limit for MP2 and CCSD(T) level using Equation 6-3 and 6-4, respectively. For the other two columns (CCSD(T)/aug-cc-pVTZ and CCSD(T)/aug-cc-pVQZ), the geometry optimizations and Gibbs free energy calculations were performed at the CCSD(T)/aug-cc-pVTZ and CCSD(T)/aug-cc-pVQZ level, respectively. The numbers shown in parenthesis are the deviations of calculated gas-phase basicities compared to the experimental values. (MAXE: maximum error; MSE: mean signed error; MUE: mean unsigned error; RMSE: root mean square error.)

kcal mol^{-1}. Only 6 gas-phase basicities (hydrogen cyanide, methanol, cyanamide, methyl hydroperoxide, acetic acid and 1,2-ethanediol) out of 41 obtained by CCSD(T)_CBS calculations fell outside the experimentally measured range. As the *ab initio* electron-correlation level increases from MP2 to CCSD(T), the accuracy gets better. From this comparison, we conclude, not unexpectedly, that accurate estimation of the electron correlation energy is important for theoretical gas-phase basicity predictions. Moreover, CCSD(T)_CBS calculations provide reliable gas-phase basicities of molecules at chemical accuracy at an affordable computational cost.

Table 6-2. Calculated and experimental gas-phase basicities of 41 small molecules (kcal mol^{-1}).*

A^-	AH	HF/ aug-cc-pVQZ	MP2/ aug-cc-pVQZ	CCSD(T) /aug-cc-pVDZ	MP2 _CBS	CCSD(T) _CBS	Exp.[183]
HO^-	water	393.7	380.0	381.9	379.8 (-3.9)	383.7 (0.0)	383.7 ±0.2
HS^-	hydrogen sulfide	346.8	342.9	343.8	342.4 (-2.5)	345.5 (+0.6)	344.9 ±1.2
CN^-	hydrogen cyanide	342.5	342.4	340.8	342.3 (-1.4)	343.1 (-0.6)	343.7 ±0.3
HC_2^-	acetylene	372.7	368.9	365.5	369.0 (-1.0)	369.5 (-0.5)	370.0 ±1.8
HO_2^-	hydrogen peroxide	375.4	368.0	367.6	367.8 (-0.8)	369.2 (+0.6)	368.6 ±0.6
HCO_2^-	formic acid	343.4	333.9	335.5	333.7 (-4.6)	336.8 (-1.5)	338.3 ±1.5
CH_3O^-	methanol	384.6	373.1	373.9	373.0 (-2.0)	375.8 (+0.8)	375.0 ±0.6
$C_2H_5O^-$	ethanol	382.0	369.3	369.9	369.2 (-2.1)	371.7 (+0.4)	371.3 ±1.1
CCl_3^-	chloroform	357.9	351.6	347.9	350.8 (+1.1)	350.5 (+0.8)	349.7 ±2.0
$NCNH^-$	cyanamide	347.5	338.7	340.7	338.5 (-5.5)	341.7 (-2.3)	344.0 ±2.0
CH_3S^-	methanethiol	354.0	348.6	349.1	348.2 (-2.4)	351.1 (+0.5)	350.6 ±2.0
$C_2H_5S^-$	ethanethiol	352.0	345.5	346.0	345.1 (-3.8)	348.1 (-0.8)	348.9 ±2.0
$CH_3CH_2CH_2O^-$	1-propanol	380.8	367.6	368.2	367.5 (-1.9)	370.0 (+0.6)	369.4 ±1.4
$(CH_3)_2CHO^-$	2-propanol	380.4	367.1	367.7	367.0 (-1.8)	369.5 (+0.7)	368.8 ±1.1
$CH_2(O)CH^-$	acetaldehyde	368.8	356.6	359.4	356.2 (-3.2)	359.7 (+0.3)	359.4 ±2.0
CH_2CN^-	acetonitrile	372.2	364.1	366.2	363.8 (-2.2)	366.3 (+0.3)	366.0 ±2.0
$CH_2NO_2^-$	nitromethane	355.5	349.5	351.3	349.0 (-1.4)	350.8 (+0.4)	350.4 ±2.0
$CH_2ClCO_2^-$	chloroacetic acid	334.5	325.3	326.4	325.0 (-3.9)	327.7 (-1.2)	328.9 ±2.0
CH_3OO^-	methyl hydroperoxide	372.1	364.6	364.2	364.4 (-3.2)	365.5 (-2.1)	367.6 ±0.7
$CH_3CH_2OO^-$	ethyl hydroperoxide	371.5	363.7	363.1	363.5 (-0.4)	364.4 (+0.5)	363.9 ±2.0
CH_3CONH^-	acetamide	365.8	354.0	354.0	353.9 (-1.1)	356.0 (+1.0)	355.0 ±2.0
$CH_3S(O)CH_2^-$	dimethyl sulfoxide	379.1	365.7	367.8	365.4 (-1.4)	368.3 (+1.5)	366.8 ±2.0
$C_6H_5S^-$	thiophenol	338.0	330.1	330.9	329.7 (-4.1)	333.3 (-0.5)	333.8 ±2.0
$CH_3C(O)CH_2^-$	acetone	373.3	360.4	363.1	360.1 (-2.1)	363.5 (+1.3)	362.2 ±2.0

Table 6-2. Continued

A^-	AH	HF/aug-cc-pVQZ	MP2/aug-cc-pVQZ	CCSD(T)/aug-cc-pVDZ	MP2_CBS	CCSD(T)_CBS	Exp.[183]
$C(CH_3)_3O^-$	*t*-butanol	379.3	365.8	366.8	365.7 (-2.2)	368.3 (+0.4)	367.9 ±1.1
$CH_3COCO_2^-$	pyruvic acid	332.2	325.0	325.7	324.8 (-1.7)	327.3 (+0.8)	326.5 ±2.8
$CF_3CO_2^-$	trifluoroacetic acid	322.9	313.8	314.8	313.6 (-3.1)	316.4 (-0.3)	316.7 ±2.0
$H_2C{=}CHCH_2O^-$	allyl alcohol	376.5	363.8	364.8	363.6 (-3.0)	366.3 (-0.3)	366.6 ±2.8
$H_2C{=}CHCO_2^-$	acrylic acid	344.0	333.9	335.2	333.7 (-3.5)	336.5 (-0.7)	337.2 ±2.8
$CH_3CH_2CO_2^-$	propanoic acid	346.7	336.6	337.7	336.4 (-4.0)	339.0 (-1.4)	340.4 ±2.0
$CH_3CO_2^-$	acetic acid	346.1	336.3	337.6	336.2 (-5.2)	338.9 (-2.5)	341.4 ±2.0
$CH_2OHCH_2O^-$	1,2-ethanediol	372.5	355.5	357.3	355.3 (-5.6)	358.4 (-2.5)	360.9 ±2.0
$CF_3CH_2O^-$	2,2,2-trifluoroethanol	362.9	352.2	352.5	352.0 (-2.1)	354.5 (+0.4)	354.1 ±2.0
$C_6H_5O^-$	phenol	350.4	339.4	340.2	339.3 (-3.6)	342.2 (-0.7)	342.9 ±1.3
$C_3H_7S^-$	1-propanethiol	351.6	345.0	345.5	344.6 (-3.3)	347.5 (-0.4)	347.9 ±2.0
$CHCl_2CO_2^-$	dichloroacetic acid	326.7	317.4	318.8	317.1 (-4.4)	320.0 (-1.5)	321.5 ±2.0
O_2^-	hydroperoxyl radical	361.8	339.9	345.2	339.7 (-7.0)	347.0 (+0.3)	346.7 ±0.8
$CH(CF_3)_2O^-$	1,1,1,3,3,3-hexafluoropropan-2-ol	344.8	334.8	334.9	334.6 (-3.8)	336.9 (-1.5)	338.4 ±2.0
$C_6H_5CO_2^-$	benzoic acid	340.3	329.7	331.2	329.4 (-3.6)	332.4 (-0.6)	333.0 ±2.0
$CH_3CH_2CHOCH_3^-$	2-butanol	379.2	365.3	366.0	365.2 (-2.3)	367.6 (+0.1)	367.5 ±2.0
$ClC_6H_4O^-$	2-chlorophenol	344.3	334.2	334.7	334.1 (-3.0)	336.9 (-0.2)	337.1 ±2.0
MAXE		15.1	-6.8	-4.5	-7.0	-2.5	
MSE		7.3	-2.6	-1.8	-2.9	-0.2	
MUE		7.4	2.7	1.9	2.9	0.8	
RMSE		8.0	3.1	2.2	3.2	1.0	

*Similar to Table 6-1, geometry optimizations and thermochemical analyses were all performed at the MP2/aug-cc-pVTZ level. The ZPVEs only include the harmonic contributions. The electronic energies on the optimized geometries were calculated at HF/aug-cc-pVQZ, MP2/aug-cc-pVQZ, CCSD(T)/aug-cc-pVDZ, and extrapolated to the complete basis set limit for MP2 and CCSD(T) using Equation 6-3 and 6-4, respectively. The numbers shown in parenthesis are the deviations of the calculated values compared to the experimental values. The deviations larger than the experimental error bars are highlighted in red.

To further check the convergence of the extrapolation approach, we chose six molecules (hydrogen cyanide, methanol, cyanamide, methyl hydroperoxide, acetic acid and 1,2-ethanediol) whose calculated gas-phase basicities deviated from the experimental values for further analysis. As shown in Equation 6-8, we computed the complete basis set limit for CCSD(T) by extrapolating the energies from CCSD(T)/aug-cc-pVTZ calculations instead of from the CCSD(T)/aug-cc-pVDZ level,

$$E_{CCSD(T)\_CBS} = E_{MP2\_CBS} + (E_{CCSD(T), aug\text{-}cc\text{-}pVTZ} - E_{MP2, aug\text{-}cc\text{-}pVTZ}) \qquad (6\text{-}8)$$

Table 6-3. The gas-phase basicity complete basis set estimations using two different extrapolation schemes. a) calculated using Equation 6-4; b) calculated using Equation 6-8. c) MP2_CBS is extrapolated from aug-cc-pVQZ and aug-cc-pV5Z energies, and HF energy is using HF/aug-cc-pV5Z. Then CCSD(T)_CBS is calculated using Equation 6-8.*

A^-	AH	a) CCSD(T)_CBS (from aug-cc-pVDZ)	b) CCSD(T)_CBS (from aug-cc-pVTZ)	c) CCSD(T)_CBS (from aug-cc-pVTZ)	Exp.[183]
CN^-	hydrogen cyanide	343.1 (-0.6)	343.2 (-0.5)	342.9 (-0.8)	343.7±0.3
CH_3O^-	methanol	375.8 (+0.8)	375.9 (+0.9)	375.7 (+0.7)	375.0±0.6
NCNH-	cyanamide	341.7 (-2.3)	341.5 (-2.5)	341.3 (-2.7)	344.0±2.0
$CH3OO^-$	methyl hydroperoxide	365.5 (-2.1)	365.6 (-2.0)	365.4 (-2.2)	367.6±0.7
$CH3CO_2^-$	acetic acid	338.9 (-2.5)	338.9 (-2.5)	338.8 (-2.6)	341.4±2.0
$CH_2OHCH_2O^-$	1,2-ethanediol	358.6 (-2.3)	358.7 (-2.2)	358.7 (-2.2)	360.9±2.0

*The numbers shown in parenthesis are the deviations of calculated gas-phase basicities compared to the experimental values. The ZPVEs only include the harmonic contributions.

As shown in Table 6-3, the CCSD(T)_CBS extrapolated from CCSD(T)/aug-cc-pVDZ and CCSD(T)/aug-cc-pVTZ levels yield almost identical gas-phase basicities. In addition, we also obtained the CBS extrapolated values of the MP2 correlation energies (E_{MP2_CBS}) from energy calculations using two larger basis sets, aug-cc-pVQZ and aug-cc-pV5Z using Equation 6-3 (where x=4 for aug-cc-pVQZ and x=5 for aug-cc-pV5Z), and the Hartree-Fock energies were taken from the results of HF/aug-cc-pV5Z calculations. As shown in Table 6-3, using the MP2 CBS energies extrapolated from larger basis sets, the gas-phase basicities obtained from CCSD(T) CBS energies have very subtle changes. Therefore, the results are likely converged, or nearly converged, for these six molecules. It indicates that the CBS limit of CCSD(T) extrapolated from CCSD(T)/aug-cc-pVDZ level is, indeed, reliable for gas-phase basicity calculations.

Following the spirit of FPA approach,[200,246] we further check the convergence of the HF, MP2 and CCSD(T) CBS limits using an extrapolation based on aug-cc-pV5Z and aug-cc-pV6Z for five representative molecules. For extrapolation of the Hartree-Fock energies, the two parameter exponential functions were used.[248,249]

$$E_X^{HF} = E_{CBS}^{HF} + a(X+1)e^{-9\sqrt{X}} \tag{6-9}$$

The MP2 and CCSD(T) CBS energies were extrapolated using Equation 6-3. As shown in Table 6-4, the gas-phase basicities calculated using MP2 energies extrapolated from smaller basis sets aug-cc-pVTZ and aug-cc-pVQZ are very close to those extrapolated gas-phase basicities using the much larger basis sets aug-cc-pV5Z and aug-cc-pV6Z. Among the five small molecules, the largest deviation of the MP2 extrapolated values is 0.39 kcal mol^{-1} for H_2O. Meanwhile, the CCSD(T) computed gas-phase basicities using the extrapolation scheme of Equation 6-4 are also very close to the CCSD(T) CBS limits. The largest deviation is also as low as 0.39 kcal mol^{-1} for C_2H_2 comparing the computed gas-phase basicities using Equation 6-4 with the CCSD(T) CBS extrapolated values based on aug-cc-pV5Z and aug-cc-pV6Z basis sets. The observed deviations from the CBS limit calculations are well below our target accuracy (1 kcal mol^{-1}). Overall, it is not currently routinely feasible to carry out MP2 and CCSD(T) calculations using aug-cc-pV5Z and aug-cc-pV6Z basis sets for molecules with more than 2 heavy atoms. Therefore, we conclude that the scheme proposed in this study provides an affordable approach for theoretical predictions of the gas-phase basicities of larger molecules within the accuracy of 1 kcal mol^{-1}.

The fact that the computed results indicate that they are likely converged suggests

Table 6-4. Calculated and experimental gas-phase basicities (ΔG in kcal mol^{-1}) of five representative small molecules. Geometry optimizations and thermochemical analyses were all performed at MP2/aug-cc-pVTZ level. The electronic energies on the optimized geometries were extrapolated to complete basis set limit for HF, MP2 and CCSD(T) level using electronic energies calculated with aug-cc-pV5Z and aug-cc-pV6Z basis sets. The numbers shown in parenthesis are the deviations of calculated gas-phase basicities compared to the experimental values. The numbers shown in bracket are the deviations of extrapolated gas-phase basicities using smaller basis sets (see text for more details) compared to the CBS estimated values using aug-cc-pV5Z and aug-cc-pV6Z basis sets (the values listed in the seventh line of each table). a) H_2O, b) H_2S, c) HCN, d) C_2H_2, e)H_2O_2

a)

	ΔG	ΔG	ΔG	MP2_CBS*	CCSD(T)_CBS**	Exp.[183]
	(RHF)	(MP2)	[CCSD(T)]			
aug-cc-pVDZ	391.56	378.08	381.91	379.84	383.67	383.7±0.2
aug-cc-pVTZ	393.37	379.85	383.74	(-3.86)	(-0.03)	
aug-cc-pVQZ	393.74	380.00	383.95	[+0.39]	[+0.06]	
aug-cc-pV5Z	393.82	379.85	383.88			
aug-cc-pV6Z	393.82	379.68	383.76			
CBS	393.82	379.45	383.61			
Δ(CBS-Exp.)	+10.12	-4.25	-0.09			

b)

	ΔG	ΔG	ΔG	MP2_CBS*	CCSD(T)_CBS**	Exp.[183]
	(RHF)	(MP2)	[CCSD(T)]			
aug-cc-pVDZ	343.51	340.72	343.81	342.37	345.46	344.9±1.2
aug-cc-pVTZ	346.15	343.00	345.39	(-2.53)	(+0.56)	
aug-cc-pVQZ	346.76	342.89	345.13	[-0.18]	[+0.33]	
aug-cc-pV5Z	347.18	342.97	345.32			
aug-cc-pV6Z	347.26	342.83	345.27			
CBS	347.28	342.55	345.13			
Δ(CBS-Exp.)	+2.38	-2.35	+0.23			

c)

	ΔG	ΔG	ΔG	MP2_CBS*	CCSD(T)_CBS**	Exp.[183]
	(RHF)	(MP2)	[CCSD(T)]			
aug-cc-pVDZ	340.17	340.04	340.80	342.30	343.06	343.7±0.3
aug-cc-pVTZ	342.43	342.45	343.30	(-1.40)	(-0.64)	
aug-cc-pVQZ	342.54	342.41	343.28	[+0.24]	[-0.04]	
aug-cc-pV5Z	342.60	342.28	343.21			
aug-cc-pV6Z	342.61	342.19	343.17			
CBS	342.61	342.06	343.10			
Δ(CBS-Exp.)	-1.09	-1.64	-0.60			

Table 6-4. Continued

d)

	ΔG	ΔG	ΔG	MP2_CBS*	CCSD(T)_CBS**	Exp.[183]
	(RHF)	(MP2)	[CCSD(T)]			
aug-cc-pVDZ	369.53	364.94	365.47	368.99	369.51	370.0±1.8
aug-cc-pVTZ	372.52	368.66	369.43	(-1.01)	(-0.49)	
aug-cc-pVQZ	372.71	368.93	369.78	[+0.09]	[-0.39]	
aug-cc-pV5Z	372.79	368.97	369.88			
aug-cc-pV6Z	372.80	368.95	369.89			
CBS	372.80	368.90	369.90			
Δ(CBS-Exp.)	+2.80	-1.10	-0.10			

e)

	ΔG	ΔG	ΔG	MP2_CBS*	CCSD(T)_CBS**	Exp.[183]
	(RHF)	(MP2)	[CCSD(T)]			
aug-cc-pVDZ	373.03	366.27	367.60	367.82	369.15	368.6±0.6
aug-cc-pVTZ	374.95	367.89	369.27	(-0.78)	(+0.55)	
aug-cc-pVQZ	375.40	368.04	369.42	[+0.35]	[+0.03]	
aug-cc-pV5Z	375.49	367.86	369.31			
aug-cc-pV6Z	375.50	367.70	-			
CBS	375.50	367.47	369.12***			
Δ(CBS-Exp.)	+6.90	-1.13	+0.52			

*The MP2_CBS energies were extrapolated based on Equation 6-3 using aug-cc-pVTZ and aug-cc-pVQZ electronic energies.
**The CCSD(T)_CBS energies were extrapolated using Equation 6-4.
***The CBS limit is extrapolated from aug-cc-pVQZ and aug-cc-pV5Z.

that the experimental values may have larger associated errors than what have been estimated. This notion is bolstered by the fact that for 35 of the cases examined we obtained results well within experimental error, while for only six cases we found more significant differences between theory and experiment. For methyl hydroperoxide, whose predicted gas-phase basicity has the largest deviation from the experimental value, we have also examined the possible rearranged species CH_2^--O-O-H and H-O-CH_2-O^- for the anion of methyl hydroperoxide, but the calculated gas-phase basicities for these two species are even poorer indicating that rearranged species are unlikely. Hence, at least for the case of methyl hydroperoxide, we suggest that it would be worthwhile reexamining the experimental value to validate that theory is failing. This is true in this case given that only one experimental measurement[250] is cited in the NIST standard reference database[183] for this compound. Further corrections examined previously, like relativistic, anharmonic effects or diagonal Born-Oppenheimer corrections are much smaller (~0.2 kcal mol^{-1})[200] than the present computed error, but given the unusual nature of this molecule we cannot rule out theoretical shortcomings entirely.

6.3.2 Anharmonicity Correction

Table 6-5. Harmonic and anharmonic ZPVEs for six molecules (H_2O_2, CH_3OH, $NCNH_2$, CH_3OOH, CH_3COOH and CH_2OHCH_2OH) and their anions computed at the MP2/aug-cc-pVTZ level. The calculated CCSD(T) CBS (using Equation 6-4) and experimental gas-phase basicities of these six molecules (in kcal mol^{-1}) are also listed. The numbers shown in parenthesis are the deviations of calculated gas-phase basicities compared to the experimental values.

molecule		Harmo-nic ZPVE (a)	Anhar-monic ZPVE (b)	b-a	ΔG [CCSD(T)] with harmonic ZPVE (c)	ΔG [CCSD(T)] with anharmonic ZPVE (d)	d-c	Exp.[183]
A^-	HO_2^-	8.35	8.21	-0.14	369.15	369.29	+0.14	368.6
AH	hydrogen peroxide	16.63	16.35	-0.28	(+0.55)	(+0.69)		±0.6
A^-	CH_3O^-	22.67	22.17	-0.50	375.78	375.77	-0.01	375.0
AH	methanol	32.55	32.06	-0.49	(+0.78)	(+0.77)		±0.6
A^-	NCNH-	12.81	12.69	-0.12	341.65	341.80	+0.15	344.0
AH	cyanamide	21.33	21.05	-0.28	(-2.35)	(-2.20)		±2.0
A^-	$CH3OO^-$	26.41	26.05	-0.36	365.46	365.60	+0.14	367.6
AH	methyl hydroperoxide	34.61	34.11	-0.50	(-2.14)	(-2.00)		±0.7
A^-	$CH3CO_2^-$	30.37	29.89	-0.48	338.86	338.90	+0.04	341.4
AH	acetic acid	39.00	38.48	-0.52	(-2.54)	(-2.50)		±2.0
A^-	$CH_2OHCH_2O^-$	44.67	43.72	-0.95	358.39	358.29	-0.10	360.9
AH	tGg’	54.12	53.31	-0.81	(-2.51)	(-2.61)		±2.0
(1,2-ethane diol)	g’Gg’	53.93	52.98	-0.95				
	gGg’	54.17	53.36	-0.81				

We further check the role anharmonic effects play on the gas-phase basicities for the molecules which were found to have relatively larger deviations from experiment. One can see from Table 6-5, the anharmonic effect lowers the ZPVE by 0.1 kcal mol^{-1} to 1.0 kcal mol^{-1}. Especially for the relatively floppy molecule 1,2-ethanediol, the anharmonic correction has the largest value of -0.95 kcal mol^{-1} among the six molecules we have examined in Table 6-5. However, the anharmonic correction is largely cancelled out when we calculate the gas-phase basicities by deducting the anharmonic correction of the molecule from its anion. As shown in Table 6-5, the anharmonic effects on the gas-phase basicities are less than or equal to 0.15 kcal mol^{-1} for all six molecules, which is much smaller than our target accuracy 1 kcal mol^{-1}. Therefore, we conclude that the harmonic ZPVE is adequate for our theoretical prediction on the gas-phase basicities.

6.3.3 Conformational Effects

For a few flexible molecules in this test set, we performed geometry optimizations from different starting geometries. Different initial conformations are usually trapped at different local minima at the end of the geometry optimization. We took the structure with the lowest free energy for the gas-phase basicity calculation when the energy difference between the two conformers was larger than 2.0 kcal mol^{-1}. Otherwise, we took the ensemble average of all low energy conformations (< 2.0 kcal mol^{-1} energy difference) based on the Maxwell-Boltzmann statistics,

$$E = \sum_i p_i \varepsilon_i \tag{6-10}$$

$$p_i = \frac{g_i e^{-\varepsilon_i / k_B T}}{\sum_i g_i e^{-\varepsilon_i / k_B T}} \tag{6-11}$$

where ε_i is the free energy of the *i*-th conformer and g_i is the degeneracy of the energy level ε_i.

To illustrate this, we carried out a conformational study on 1,2-ethanediol. As shown in Figure 6-1a to 6-1d, four different local minima (tTt, tGg’, gGg’ and g’Gg’) were found for 1,2-ethanediol at the MP2/aug-cc-pVTZ level, which is consistent with previous studies.[251-253] The conformer tGg’ with a weak intramolecular hydrogen bond is 2.0 kcal mol^{-1} lower in total free energy than the conformer tTt without the intramolecular hydrogen bond. The other two conformers gGg’ and g’Gg’ are 0.5 kcal mol^{-1} and 0.3 kcal mol^{-1} higher than the conformer tGg’, respectively. Previous study has shown that the conformer gGg’ has a lower free energy than g’Gg’ based on MP2/6-31G* calculations using the geometries optimized at the HF/6-31G* level[251], while in this study, we find g’Gg’ is more stable than gGg’ at the MP2/aug-cc-pVTZ level. Moreover, for the anion of 1,2-ethanediol ($CH_2OHCH_2O^-$), the conformer shown in Figure 6-1f has a stronger intramolecular hydrogen bonding interaction in terms of the donor-acceptor distance. Compared to the neutral 1,2-ethanediol at the tGg’ configuration, the distance between hydrogen-donor and oxygen-acceptor is decreased from 2.32 Å to 1.63 Å, and the O-H-O angle is increased from 108.7^{o} to 137.0^{o}, and thus the total free energy of the conformer shown in Figure 6-1f is 12.2 kcal mol^{-1} lower than the conformer without the intramolecular hydrogen bond shown in Figure 6-1e. The gas-phase basicity calculations on 1,2-ethanediol further confirm that the structures with the intramolecular hydrogen bonds should be used for computing chemical properties. One can also see from Table 6-6, the calculated CCSD(T)_CBS gas-phase basicity of 1,2-ethanediol has a 2.5 kcal mol^{-1} deviation from experiment using the geometries with the lower energies (conformer f and ensemble average over b, c and d). On the other hand, the CCSD(T)_CBS predicted value derived from conformer e) and a) (see Figure 6-1) has a

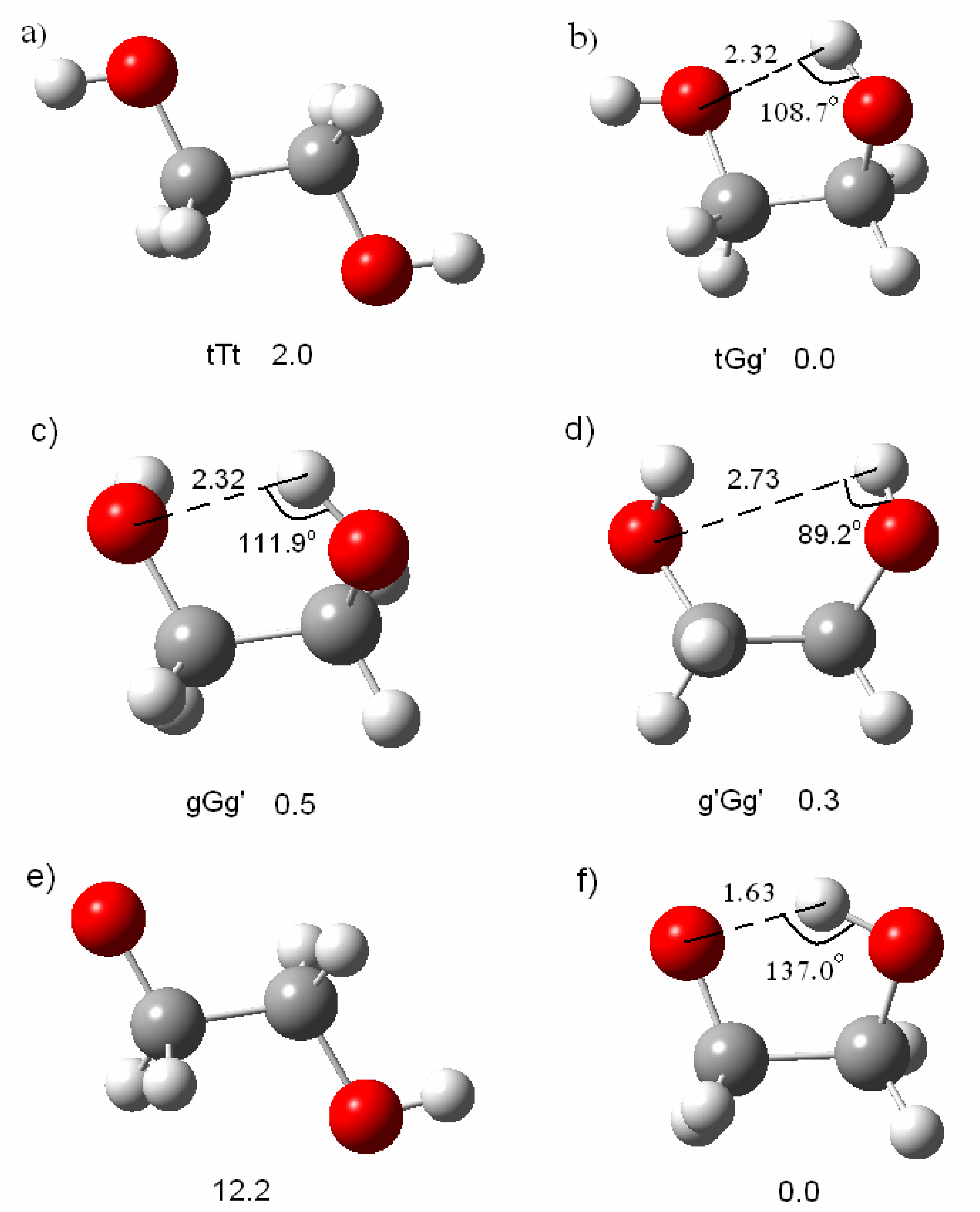


Figure 6-1. Different local minima for 1,2-ethanediol CH_2OHCH_2OH (a, b, c and d) and for the anion of 1,2-ethanediol $CH_2OHCH_2O^-$ (e and f) optimized at the MP2/aug-cc-pVTZ level. The number below each conformer is the relative free energy in kcal mol^{-1}. (Carbon, oxygen and hydrogen atoms are represented in gray, red and white color, respectively. The distance between the oxygen atom and hydrogen atom is in Å.) The ZPVEs only include the harmonic contributions.

larger deviation of 7.7 kcal mol^{-1}. This shows that conformational effects are relevant for

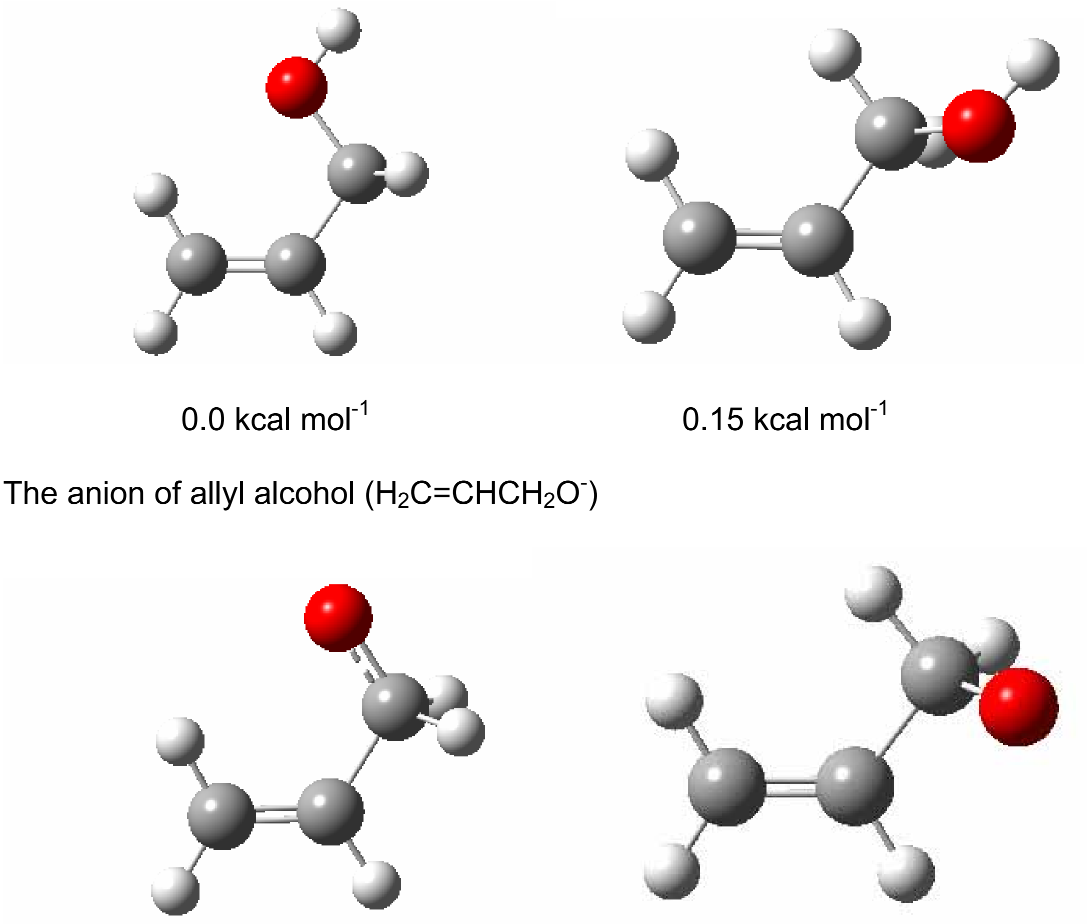
Table 6-6. The gas-phase basicity of the anion of 1,2-ethanediol calculated using different local minima.

A^-	AH	MP2_CBS	CCSD(T)_CBS	Exp.
e)	a)	366.0 (+5.1)	368.6 (+7.7)	360.9±2.0
f)	b), c), d)*	355.3 (-5.6)	358.4 (-2.5)	

*Ensemble average over conformers b), c) and d).

theoretical predictions of the gas-phase basicities of molecules. Thus sampling represents yet another challenge associated with computing gas-phase basicities using extraordinarily sophisticated computational techniques.[200] Further conformational

1) allyl alcohol (H_2C=$CHCH_2OH$)



0.0 kcal mol^{-1}

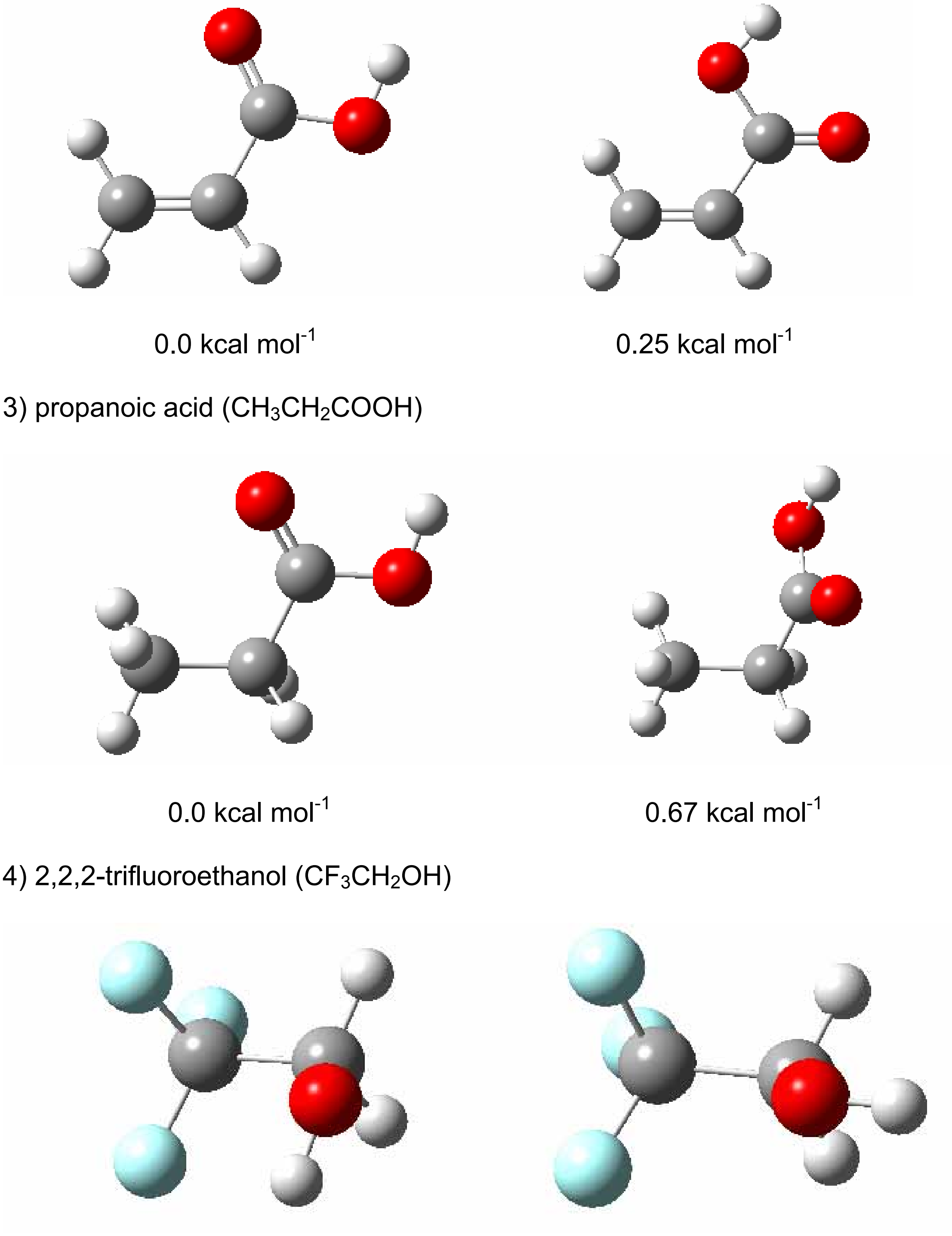
0.15 kcal mol^{-1}

The anion of allyl alcohol (H_2C=$CHCH_2O^-$)

0.0 kcal mol^{-1}

1.57 kcal mol^{-1}

Figure 6-2. Structures optimized at the MP2/aug-cc-pVTZ level. The relative total free energies are given below each structure correspondingly.



2) acrylic acid ($H_2C{=}CHCOOH$)

0.0 kcal mol^{-1}

0.25 kcal mol^{-1}

3) propanoic acid (CH_3CH_2COOH)

0.0 kcal mol^{-1}

0.67 kcal mol^{-1}

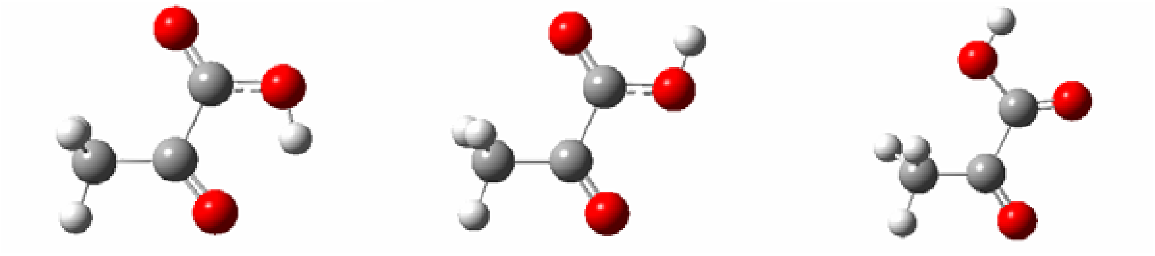
4) 2,2,2-trifluoroethanol (CF_3CH_2OH)

0.0 kcal mol^{-1}

1.16 kcal mol^{-1}

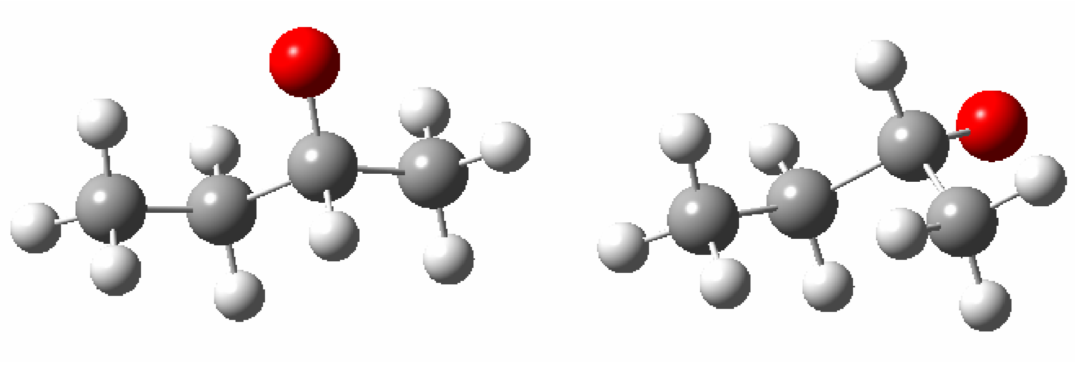
5) pyruvic acid ($CH_3COCOOH$)

Figure 6-2. Continued



0.0 kcal mol^{-1} 2.15 kcal mol^{-1} 3.38 kcal mol^{-1}

6) the anion of 2-butanol ($CH_3CH_2CHOCH_3^-$)



0.0 kcal mol^{-1} 1.79 kcal mol^{-1}

Figure 6-2. Continued

studies for allyl alcohol, acrylic acid, propanoic acid, 2,2,2-trifluoroethanol, pyruvic acid and 2-butanol are presented in Figure 6-2.

6.4 Conclusions

Through the theoretical study of the gas-phase basicities of 41 small molecules, chemical accuracy was achieved via CCSD(T) calculations with CBS extrapolation. For 35 of the cases studied theory and experiment were in excellent accord, while for six cases (hydrogen cyanide, methanol, cyanamide, methyl hydroperoxide, acetic acid and 1,2-ethanediol) theory predicted values outside of the experimental error bars. We suggested that a re-examination of the experimental value for methyl hydroperoxide will help us determine whether some aspect of the theoretical approach is less than optimal or if the experimental uncertainties are larger than currently believed. The electron correlation energy was found to be an important component in the theoretical estimation of gas-phase basicities. The least inexpensive *ab initio* electron correlation method MP2, which scales with the fifth power of molecular size, was not adequate for gas-phase basicity prediction. For cases, where experimental gas-phase basicities are not available, or large uncertainties (~3.0 kcal mol^{-1}) are associated with the available values, the computational procedure proposed in this study provides a validated approach to accurately predict the gas-phase basicities of molecules with near chemical accuracy. Even though the computational expense scales with the seventh power of the molecular size for CCSD(T) calculations, modern parallel implementation of CCSD(T) calculations[254-258] and low-order scaling local electron correlation methods[259-262] have extended the power of coupled-cluster theory to systems beyond 10 heavy atoms.

CHAPTER 7
CONCLUSIONS

In this dissertation, the implementation of the divide-and-conquer (DC) algorithm, an algorithm with the potential to aid the achievement of true linear scaling within Hartree-Fock (HF) and second-order Møller-Plesset perturbation (MP2) theories are revisited. The DC algorithm for HF calculations was validated on polyglycines, polyalanines and eleven real three-dimensional proteins of up to 608 atoms in this thesis. We also found that a fragment-based initial guess using molecular fractionation with conjugated caps (MFCC) method significantly reduces the number of SCF cycles and even is capable of achieving convergence for some globular proteins where the simple superposition of atomic densities (SAD) initial guess fails. For DC-MP2 calculations, after localized molecular orbitals (LMO) of each subsystem are obtained from the DC-HF calculations, the correlation energy of the whole system can be derived by taking the sum of the local electron correlation of each subsystem. Preliminary DC-MP2 results on extended polyglycine systems show the linear-scaling behavior.

The AF-QM/MM method shows good agreement with standard self-consistent field (SCF) calculations of the NMR chemical shieldings for the mini-protein Trp-cage. The root mean square errors (RMSEs) for ^{1}H, ^{13}C and ^{15}N NMR chemical shieldings are equal to or less than 0.09ppm, 0.32ppm, and 0.78ppm, respectively, for all Hartree-Fock (HF) and density functional theory (DFT) calculations reported in this thesis. The environmental electrostatic potential is necessary to accurately reproduce the NMR chemical shieldings using the AF-QM/MM approach. The point charge models provided by AMBER, AM1/CM2, PM3/CM1 and PM3/CM2 all effectively model the electrostatic field. The latter three point charge models are generated via semiempirical linear-scaling SCF calculations of the entire protein system. The correlations between experimental ^{1}H NMR chemical shifts and theoretical predictions are >0.95 for AF-QM/MM calculations using B3LYP with the 6-31G**, 6-311G** and 6-311++G** basis sets. Our study, not unexpectedly, finds that conformational changes within a protein structure play an important role in the accurate prediction of the experimental NMR chemical shifts from theory.

In the study of *ab initio* protein folding, we have shown the sum of the HF energy and force field (LJ6) derived dispersion energy (HF + LJ6) is well correlated with the energies obtained using second-order MP2 theory. Furthermore, when we randomly scrambled the Lennard-Jones parameters, the correlation between the MP2 energy and the sum of HF energy and dispersive energy (HF+LJ6) significantly drops, which indicates that the choice of Lennard-Jones parameters is important.

The overall accuracy for different *ab initio* methods to calculate the molecular gas-phase basicities are compared and the accuracy in descending order is CCSD(T)_CBS > CCSD(T)/aug-cc-pVDZ > (MP2/aug-cc-pVQZ$\approx$MP2_CBS) > HF/ aug-cc-pVQZ. The best root-mean-squared-error obtained was 1.0 kcal mol^{-1} at the CCSD(T)_CBS//MP2/aug-cc-pVTZ level for a test set of 41 small molecules. Clearly, accurate calculations for the electron correlation energy are important for the theoretical prediction of molecular gas-phase basicities. However, conformational effects were also found to be relevant in several instances when more complicated molecules were examined.

LIST OF REFERENCES

(1) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry : introduction to advanced electronic structure theory*, 1st. ed.; McGraw-Hill: New York, 1989.

(2) Parr, R. G.; Yang, W. T. *Annual Review of Physical Chemistry* **1995**, *46*, 701.

(3) Bartlett, R. J.; Musial, M. *Reviews of Modern Physics* **2007**, *79*, 291.

(4) Strout, D. L.; Scuseria, G. E. *Journal of Chemical Physics* **1995**, *102*, 8448.

(5) Schwegler, E.; Challacombe, M. *Journal of Chemical Physics* **1996**, *105*, 2726.

(6) Goedecker, S. *Reviews of Modern Physics* **1999**, *71*, 1085.

(7) Fedorov, D. G.; Kitaura, K. *Journal of Physical Chemistry A* **2007**, *111*, 6904.

(8) Challacombe, M.; Schwegler, E. *Journal of Chemical Physics* **1997**, *106*, 5526.

(9) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. X. *Journal of Physical Chemistry A* **1999**, *103*, 1913.

(10) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chemical Physics Letters* **1994**, *230*, 8.

(11) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. *Chemical Physics Letters* **1996**, *253*, 268.

(12) Scuseria, G. E. *Journal of Physical Chemistry A* **1999**, *103*, 4782.

(13) Korchowiec, J.; Lewandowski, J.; Makowski, M.; Gu, F. L.; Aoki, Y. *Journal of Computational Chemistry* **2009**, *30*, 2515.

(14) Jiang, N.; Ma, J.; Jiang, Y. S. *Journal of Chemical Physics* **2006**, *124*, 114112.

(15) Daniels, A. D.; Scuseria, G. E. *Journal of Chemical Physics* **1999**, *110*, 1321.

(16) Yang, W. T. *Physical Review Letters* **1991**, *66*, 1438.

(17) Yang, W. T.; Lee, T. S. *Journal of Chemical Physics* **1995**, *103*, 5674.

(18) Dixon, S. L.; Merz, K. M. *Journal of Chemical Physics* **1996**, *104*, 6643.

(19) Dixon, S. L.; Merz, K. M. *Journal of Chemical Physics* **1997**, *107*, 879.

(20) Kobayashi, M.; Imamura, Y.; Nakai, H. *Journal of Chemical Physics* **2007**, *127*, 074103.

(21) Kobayashi, M.; Nakai, H. *Journal of Chemical Physics* **2008**, *129*, 044103.

(22) Shaw, D. M.; St-Amant, A. *Journal of Theoretical & Computational Chemistry* **2004**, *3*, 419.

(23) He, X.; Wang, B.; Merz, K. M. *Journal of Physical Chemistry B* **2009**, *113*, 10380.

(24) He, X.; Fusti-Molnar, L.; Cui, G. L.; Merz, K. M. *Journal of Physical Chemistry B* **2009**, *113*, 5290.

(25) Park, B.; Levitt, M. *Journal of Molecular Biology* **1996**, *258*, 367.

(26) Lazaridis, T.; Karplus, M. *Current Opinion in Structural Biology* **2000**, *10*, 139.

(27) Lazaridis, T.; Karplus, M. *Journal of Molecular Biology* **1999**, *288*, 477.

(28) Dominy, B. N.; Brooks, C. L. *Journal of Computational Chemistry* **2002**, *23*, 147.

(29) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins-Structure Function and Genetics* **2002**, *48*, 404.

(30) Wollacott, A. M.; Merz, K. M. *Journal of Chemical Theory and Computation* **2007**, *3*, 1609.

(31) Lee, M. R.; Kollman, P. A. *Structure* **2001**, *9*, 905.

(32) Vondrasek, J.; Bendova, L.; Klusak, V.; Hobza, P. *Journal of the American Chemical Society* **2005**, *127*, 2615.

(33) Brändén, C.-I.; Tooze, J. *Introduction to protein structure*, 2nd ed.; Garland Pub.: New York, 1999.

(34) He, X.; Fusti-Molnar, L.; Merz, K. M. *Journal of Physical Chemistry A* **2009**, *113*, 10096.

(35) He, X.; Ayers, K.; Brothers, E.; Merz, K. M.; QUICK; University of Florida: Gainesville; FL; 2008.

(36) Obara, S.; Saika, A. *Journal of Chemical Physics* **1986**, *84*, 3963.

(37) Headgordon, M.; Pople, J. A. *Journal of Chemical Physics* **1988**, *89*, 5777.

(38) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *Journal of Computational Chemistry* **1993**, *14*, 1347.

(39) Headgordon, M.; Pople, J. A.; Frisch, M. J. *Chemical Physics Letters* **1988**, *153*, 503.

(40) Schwegler, E.; Challacombe, M. *Journal of Chemical Physics* **1999**, *111*, 6223.

(41) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chemical Physics Letters* **1996**, *248*, 43.

(42) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.

(43) Shao, Y. H.; White, C. A.; Head-Gordon, M. *Journal of Chemical Physics* **2001**, *114*, 6572.

(44) Ochsenfeld, C. *Chemical Physics Letters* **2000**, *327*, 216.

(45) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. *Journal of Chemical Physics* **1998**, *109*, 1663.

(46) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chemical Physics Letters* **2002**, *351*, 475.

(47) Fedorov, D. G.; Kitaura, K. *Chemical Physics Letters* **2006**, *433*, 182.

(48) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagase, S. *Journal of Computational Chemistry* **2007**, *28*, 1476.

(49) He, X.; Zhang, J. Z. H. *Journal of Chemical Physics* **2005**, *122*, 031103.

(50) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chemical Physics Letters* **2000**, *318*, 614.

(51) Zhang, D. W.; Zhang, J. Z. H. *Journal of Theoretical & Computational Chemistry* **2004**, *3*, 43.

(52) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *Journal of Physical Chemistry B* **2003**, *107*, 12039.

(53) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *Journal of Chemical Physics* **2004**, *120*, 1145.

(54) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Journal of Computational Chemistry* **2004**, *25*, 1431.

(55) Zhang, D. W.; Zhang, J. Z. H. *Journal of Chemical Physics* **2003**, *119*, 3599.

(56) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. K. *Chemical Physics Letters* **2004**, *394*, 293.

(57) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. *Journal of Chemical Physics* **2004**, *120*, 839.

(58) Becke, A. D. *Journal of Chemical Physics* **1988**, *88*, 2547.

(59) He, X.; Zhang, J. Z. H. *Journal of Chemical Physics* **2006**, *124*, 184703.

(60) Exner, T. E.; Mezey, P. G. *Journal of Physical Chemistry A* **2004**, *108*, 4301.

(61) Li, W.; Li, S. H.; Jiang, Y. S. *Journal of Physical Chemistry A* **2007**, *111*, 2193.

(62) Gauss, J. *Journal of Chemical Physics* **1993**, *99*, 3629.

(63) Salter, E. A.; Trucks, G. W.; Bartlett, R. J. *Journal of Chemical Physics* **1989**, *90*, 1752.

(64) Cui, Q.; Karplus, M. *Journal of Physical Chemistry B* **2000**, *104*, 3721.

(65) Challacombe, M.; Schwegler, E.; White, C.; Johnson, B.; Gill, P.; HeadGordon, M. *Abstracts of Papers of the American Chemical Society* **1997**, *213*, 57.

(66) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Headgordon, M. *Chemical Physics Letters* **1994**, *230*, 8.

(67) White, C. A.; Johnson, B. G.; Gill, P. M. W.; HeadGordon, M. *Chemical Physics Letters* **1996**, *253*, 268.

(68) Kohn, W. *Physical Review Letters* **1996**, *76*, 3168.

(69) Exner, T. E.; Mezey, P. G. *Journal of Physical Chemistry A* **2002**, *106*, 11791.

(70) Fusti-Molnar, L. *Journal of Chemical Physics* **2003**, *119*, 11080.

(71) Fusti-Molnar, L.; Pulay, P. *Journal of Chemical Physics* **2002**, *117*, 7827.

(72) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L.; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. *Physical Chemistry Chemical Physics* **2006**, *8*, 3172.

(73) Gogonea, V.; Westerhoff, L. M.; Merz, K. M. *Journal of Chemical Physics* **2000**, *113*, 5604.

(74) Kobayashi, M.; Nakai, H. *International Journal of Quantum Chemistry* **2009**, *109*, 2227.

(75) Akama, T.; Fujii, A.; Kobayashi, M.; Nakai, H. *Molecular Physics* **2007**, *105*, 2799.

(76) Akama, T.; Kobayashi, M.; Nakai, H. *Journal of Computational Chemistry* **2007**, *28*, 2003.

(77) Kobayashi, M.; Akama, T.; Nakai, H. *Journal of Chemical Physics* **2006**, *125*, 204106.

(78) Exner, T. E.; Mezey, P. G. *Journal of Computational Chemistry* **2003**, *24*, 1980.

(79) Exner, T. E.; Mezey, P. G. *Physical Chemistry Chemical Physics* **2005**, *7*, 4061.

(80) Fedorov, D. G.; Kitaura, K. *Journal of Chemical Physics* **2005**, *123*, 134103.

(81) Chen, X. H.; Zhang, J. Z. H. *Journal of Chemical Physics* **2006**, *125*, 044903.

(82) Chen, X. H.; Zhang, Y. K.; Zhang, J. Z. H. *Journal of Chemical Physics* **2005**, *122*, 184105.

(83) Wüthrich, K. *NMR of proteins and nucleic acids*; Wiley: New York, 1986.

(84) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128.

(85) Robustelli, P.; Cavalli, A.; Vendruscolo, M. *Structure* **2008**, *16*, 1764.

(86) Spera, S.; Bax, A. *Journal of the American Chemical Society* **1991**, *113*, 5490.

(87) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 9615.

(88) Meiler, J.; Baker, D. *Proceedings of the National Academy of Sciences of the United States of America* **2003**, *100*, 15404.

(89) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G. H.; Eletsky, A.; Wu, Y. B.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 4685.

(90) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *Journal of Biomolecular Nmr* **2009**, *43*, 63.

(91) Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. *Journal of Biomolecular Nmr* **2003**, *26*, 215.

(92) Xu, X. P.; Case, D. A. *Journal of Biomolecular Nmr* **2001**, *21*, 321.

(93) Wang, B.; Brothers, E. N.; van der Vaart, A.; Merz, K. M. *Journal of Chemical Physics* **2004**, *120*, 11392.

(94) Wang, B.; Merz, K. M. *Journal of Chemical Theory and Computation* **2006**, *2*, 209.

(95) Wolinski, K.; Hinton, J. F.; Pulay, P. *Journal of the American Chemical Society* **1990**, *112*, 8251.

(96) Oldfield, E. *Philosophical Transactions of the Royal Society B-Biological Sciences* **2005**, *360*, 1347.

(97) Helgaker, T.; Jaszunski, M.; Ruud, K. *Chemical Reviews* **1999**, *99*, 293.

(98) Haser, M.; Ahlrichs, R.; Baron, H. P.; Weis, P.; Horn, H. *Theoretica Chimica Acta* **1992**, *83*, 455.

(99) Ochsenfeld, C.; Kussmann, J.; Koziol, F. *Angewandte Chemie-International Edition* **2004**, *43*, 4485.

(100) Kussmann, J.; Ochsenfeld, C. *Journal of Chemical Physics* **2007**, *127*, 054103.

(101) Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S. *Chemical Physics Letters* **2007**, *445*, 331.

(102) Xie, W. S.; Song, L. C.; Truhlar, D. G.; Gao, J. L. *Journal of Physical Chemistry B* **2008**, *112*, 14124.

(103) Xie, W. S.; Gao, J. L. *Journal of Chemical Theory and Computation* **2007**, *3*, 1890.

(104) Maseras, F.; Morokuma, K. *Journal of Computational Chemistry* **1995**, *16*, 1170.

(105) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *Journal of Physical Chemistry* **1996**, *100*, 19357.

(106) Dedios, A. C.; Oldfield, E. *Chemical Physics Letters* **1993**, *205*, 108.

(107) Scheurer, C.; Skrynnikov, N. R.; Lienin, S. F.; Straus, S. K.; Bruschweiler, R.; Ernst, R. R. *Journal of the American Chemical Society* **1999**, *121*, 4242.

(108) Ditchfield, R. *Molecular Physics* **1974**, *27*, 789.

(109) London, F. *Journal De Physique Et Le Radium* **1937**, *8*, 397.

(110) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe,

M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; *Gaussian 03* revision D.01. Gaussian Inc. Wallingford CT., 2004

(111) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nature Structural Biology* **2002**, *9*, 425.

(112) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926.

(113) Nijboer, B. R. A.; Ruijgrok, T. W. *Journal of Statistical Physics* **1988**, *53*, 361.

(114) Darden, T.; Pearlman, D.; Pedersen, L. G. *Journal of Chemical Physics* **1998**, *109*, 10921.

(115) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327.

(116) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *Journal of Chemical Physics* **1984**, *81*, 3684.

(117) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *Journal of Computational Chemistry* **2005**, *26*, 1668.

(118) Mulliken, R. S. *Journal of Chemical Physics* **1955**, *23*, 1841.

(119) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *Journal of Chemical Physics* **1985**, *83*, 735.

(120) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *Journal of Computer-Aided Molecular Design* **1995**, *9*, 87.

(121) Zhu, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Journal of Chemical Physics* **1998**, *109*, 9117.

(122) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *Journal of the American Chemical Society* **1985**, *107*, 3902.

(123) Stewart, J. J. P. *Journal of Computational Chemistry* **1989**, *10*, 209.

(124) Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, J. J.; Brothers, E. N.; Suarez, D.; Westerhoff, L. M.; Merz, K. M. J.; *DivCon* (The Pennsylvania State University, U. P., PA,1999).

(125) Markwick, P. R. L.; Sprangers, R.; Sattler, M. *Journal of the American Chemical Society* **2007**, *129*, 8048.

(126) Burgi, R.; Pitera, J.; van Gunsteren, W. F. *Journal of Biomolecular Nmr* **2001**, *19*, 305.

(127) Case, D. A.; Scheurer, C.; Bruschweiler, R. *Journal of the American Chemical Society* **2000**, *122*, 10390.

(128) Hoch, J. C.; Dobson, C. M.; Karplus, M. *Biochemistry* **1982**, *21*, 1118.

(129) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *Journal of the American Chemical Society* **2002**, *124*, 11258.

(130) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins-Structure Function and Bioinformatics* **2006**, *65*, 712.

(131) Altona, C.; Sundaral.M. *Journal of the American Chemical Society* **1972**, *94*, 8205.

(132) Muller-Dethlefs, K.; Hobza, P. *Chemical Reviews* **2000**, *100*, 143.

(133) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Biochemistry* **1994**, *33*, 10026.

(134) Fersht, A. R. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 1525.

(135) Riley, K. E.; Merz, K. M. *Journal of Physical Chemistry B* **2006**, *110*, 15650.

(136) Nakanishi, I.; Fedorov, D. G.; Kitaura, K. *Proteins-Structure Function and Bioinformatics* **2007**, *68*, 145.

(137) Hobza, P.; Sponer, J. *Chemical Reviews* **1999**, *99*, 3247.

(138) Hobza, P.; Sponer, J.; Polasek, M. *Journal of the American Chemical Society* **1995**, *117*, 792.

(139) Hobza, P.; Sponer, J. *Chemical Physics Letters* **1998**, *288*, 7.

(140) Cybulski, S. M.; Chalasinski, G.; Moszynski, R. *Journal of Chemical Physics* **1990**, *92*, 4357.

(141) Chalasinski, G.; Szczesniak, M. M. *Molecular Physics* **1988**, *63*, 205.

(142) Cybulski, S. M.; Bledson, T. M.; Toczylowski, R. R. *Journal of Chemical Physics* **2002**, *116*, 11039.

(143) Tomasi, J.; Mennucci, B.; Cammi, R. *Chemical Reviews* **2005**, *105*, 2999.

(144) Fedorov, D. G.; Kitaura, K.; Li, H.; Jensen, J. H.; Gordon, M. S. *Journal of Computational Chemistry* **2006**, *27*, 976.

(145) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *Journal of Computational Chemistry* **2003**, *24*, 669.

(146) Li, H.; Jensen, J. H. *Journal of Computational Chemistry* **2004**, *25*, 1449.

(147) Barone, V.; Cossi, M.; Tomasi, J. *Journal of Chemical Physics* **1997**, *107*, 3210.

(148) Hobza, P.; Selzle, H. L.; Schlag, E. W. *Journal of Physical Chemistry* **1996**, *100*, 18790.

(149) Bonneau, R.; Strauss, C. E. M.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmstrom, L.; Robertson, T.; Baker, D. *Journal of Molecular Biology* **2002**, *322*, 65.

(150) Feig, M.; Karanicolas, J.; Brooks, C. L. *Journal of Molecular Graphics & Modelling* **2004**, *22*, 377.

(151) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. *Proteins-Structure Function and Genetics* **2003**, *53*, 76.

(152) Wollacott, A. M.; Merz, K. M. *Journal of Chemical Theory and Computation* **2006**, *2*, 1070.

(153) Dunning, T. H. *Journal of Physical Chemistry A* **2000**, *104*, 9062.

(154) Boys, S. F.; Bernardi, F. *Molecular Physics* **1970**, *19*, 553.

(155) Wang, J. M.; Cieplak, P.; Kollman, P. A. *Journal of Computational Chemistry* **2000**, *21*, 1049.

(156) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157.

(157) Johnson, E. R.; Becke, A. D. *Journal of Chemical Physics* **2005**, *123*, 024101.

(158) Becke, A. D.; Johnson, E. R. *Journal of Chemical Physics* **2006**, *124*, 014104.

(159) Tuttle, T.; Thiel, W. *Physical Chemistry Chemical Physics* **2008**, *10*, 2159.

(160) Gonzalez, C.; Lim, E. C. *Journal of Physical Chemistry A* **2003**, *107*, 10105.

(161) Ahlrichs, R.; Penco, R.; Scoles, G. *Chemical Physics* **1977**, *19*, 119.

(162) Becke, A. D.; Johnson, E. R. *Journal of Chemical Physics* **2005**, *123*, 154101.

(163) Grimme, S. *Journal of Computational Chemistry* **2004**, *25*, 1463.

(164) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Physical Review Letters* **2004**, *93*, 153004.

(165) Besler, B. H.; Merz, K. M.; Kollman, P. A. *Journal of Computational Chemistry* **1990**, *11*, 431.

(166) Jakalian, A.; Jack, D. B.; Bayly, C. I. *Journal of Computational Chemistry* **2002**, *23*, 1623.

(167) Shibasaki, K.; Fujii, A.; Mikami, N.; Tsuzuki, S. *Journal of Physical Chemistry A* **2006**, *110*, 4397.

(168) Molnar, L. F.; He, X.; Wang, B.; Merz, K. M. *Journal of Chemical Physics* **2009**, *131*, 065102.

(169) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophysical Journal* **2008**, *94*, L75.

(170) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2005**, *1*, 1133.

(171) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2007**, *3*, 2011.

(172) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Computer Physics Communications* **2005**, *167*, 103.

(173) Zhao, Y.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2007**, *3*, 289.

(174) Zhao, Y.; Truhlar, D. G. *Accounts of Chemical Research* **2008**, *41*, 157.

(175) Leach, A. R.; *Molecular Modeling Principles and Practice* 2nd Ed. Prentice-Hall. **2001**.

(176) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry B* **2007**, *111*, 408.

(177) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry B* **2006**, *110*, 16066.

(178) Pliego, J. R.; Riveros, J. M. *Physical Chemistry Chemical Physics* **2002**, *4*, 1622.

(179) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *Journal of Physical Chemistry A* **1998**, *102*, 7787.

(180) Zhan, C. G.; Dixon, D. A. *Journal of Physical Chemistry A* **2001**, *105*, 11534.

(181) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *Journal of Physical Chemistry A* **2004**, *108*, 6532.

(182) Pliego, J. R.; Riveros, J. M. *Chemical Physics Letters* **2000**, *332*, 597.

(183) Lias, S. G.; Bartness, J. E.; Liebman, J. F.; Holmes, J. L.; Levin, R. D.; Mallard, W. G.; Ion Energetics Data. In *NIST Chemistry WebBook NIST Standard Reference Database Number 69*; Linstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Technology: Gaithersburg, MD, March 2003.

(184) Meot-Ner, M. *International Journal of Mass Spectrometry* **2003**, *227*, 525.

(185) Szulejko, J. E.; Mcmahon, T. B. *Journal of the American Chemical Society* **1993**, *115*, 7839.

(186) Hunter, E. P. L.; Lias, S. G. *Journal of Physical and Chemical Reference Data* **1998**, *27*, 413.

(187) Lias, S. G.; Liebman, J. F.; Levin, R. D. *Journal of Physical and Chemical Reference Data* **1984**, *13*, 695.

(188) Pople, J. A. *Angewandte Chemie-International Edition* **1999**, *38*, 1894.

(189) Curtiss, L. A.; Redfern, P. C.; Frurip, D. J.; Lipkowitz, K. B.; Boyd, D. B.; *Reviews in Computational Chemistry* Wiley-VCH New York, vol. 15, p. 147.

(190) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Headgordon, M. *Chemical Physics Letters* **1989**, *157*, 479.

(191) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chemical Physics Letters* **1998**, *286*, 243.

(192) Martin, J. M. L. *Chemical Physics Letters* **1996**, *259*, 669.

(193) Ervin, K. M.; DeTuro, V. F. *Journal of Physical Chemistry A* **2002**, *106*, 9947.

(194) Koppel, I. A.; Burk, P.; Koppel, I.; Leito, I.; Sonoda, T.; Mishima, M. *Journal of the American Chemical Society* **2000**, *122*, 5114.

(195) Smith, B. J.; Radom, L. *Chemical Physics Letters* **1995**, *245*, 123.

(196) Smith, B. J.; Radom, L. *Chemical Physics Letters* **1994**, *231*, 345.

(197) Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Physical Chemistry Chemical Physics* **2005**, *7*, 3070.

(198) Range, K.; Lopez, C. S.; Moser, A.; York, D. M. *Journal of Physical Chemistry A* **2006**, *110*, 791.

(199) Burk, P.; Koppel, I. A.; Koppel, I.; Leito, I.; Travnikova, O. *Chemical Physics Letters* **2000**, *323*, 482.

(200) Czakó, G.; Mátyus, E.; Simmonett, A. C.; Császár, A. G.; Schaefer, H. F.; Allen, W. D. *Journal of Chemical Theory and Computation* **2008**, *4*, 1220.

(201) Burk, P.; Tamp, S. *Journal of Molecular Structure-Theochem* **2003**, *638*, 119.

(202) Ewing, N. P.; Pallante, G. A.; Zhang, X.; Cassady, C. J. *Journal of Mass Spectrometry* **2001**, *36*, 875.

(203) Gal, J. F.; Maria, P. C.; Raczynska, E. D. *Journal of Mass Spectrometry* **2001**, *36*, 699.

(204) Deakyne, C. A. *International Journal of Mass Spectrometry* **2003**, *227*, 601.

(205) Parthiban, S.; Martin, J. M. L. *Journal of Chemical Physics* **2001**, *115*, 2051.

(206) Tsushima, S.; Yang, T. X.; Suzuki, A. *Chemical Physics Letters* **2001**, *334*, 365.

(207) Martin, J. M. L.; de Oliveira, G. *Journal of Chemical Physics* **1999**, *111*, 1843.

(208) Ochterski, J. W.; Petersson, G. A.; Montgomery, J. A. *Journal of Chemical Physics* **1996**, *104*, 2598.

(209) Montgomery, J. A.; Ochterski, J. W.; Petersson, G. A. *Journal of Chemical Physics* **1994**, *101*, 5900.

(210) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *Journal of Chemical Physics* **1999**, *110*, 2822.

(211) Smith, B. J.; Radom, L. *Journal of the American Chemical Society* **1993**, *115*, 4885.

(212) Ruscic, B.; Boggs, J. E.; Burcat, A.; Császár, A. G.; Demaison, J.; Janoschek, R.; Martin, J. M. L.; Morton, M. L.; Rossi, M. J.; Stanton, J. F.; Szalay, P. G.; Westmoreland, P. R.; Zabel, F.; Bérces, T. *Journal of Physical and Chemical Reference Data* **2005**, *34*, 573.

(213) Karton, A.; Rabinovich, E.; Martin, J. M. L.; Ruscic, B. *Journal of Chemical Physics* **2006**, *125*, 144108.

(214) Császár, A. G.; Leininger, M. L.; Szalay, V. *Journal of Chemical Physics* **2003**, *118*, 10631.

(215) Feller, D.; Peterson, K. A.; de Jong, W. A.; Dixon, D. A. *Journal of Chemical Physics* **2003**, *118*, 3510.

(216) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. *Journal of Chemical Physics* **2004**, *121*, 11599.

(217) Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J. M. L.; Kállay, M.; Gauss, J. *Journal of Chemical Physics* **2004**, *120*, 4129.

(218) Bomble, Y. J.; Vázquez, J.; Kállay, M.; Michauk, C.; Szalay, P. G.; Császár, A. G.; Gauss, J.; Stanton, J. F. *Journal of Chemical Physics* **2006**, *125*, 064108.

(219) Harding, M. E.; Vázquez, J.; Ruscic, B.; Wilson, A. K.; Gauss, J.; Stanton, J. F. *Journal of Chemical Physics* **2008**, *128*, 114111.

(220) East, A. L. L.; Allen, W. D. *Journal of Chemical Physics* **1993**, *99*, 4638.

(221) Gonzales, J. M.; Pak, C.; Cox, R. S.; Allen, W. D.; Schaefer, H. F.; Császár, A. G.; Tarczay, G. *Chemistry-a European Journal* **2003**, *9*, 2173.

(222) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. *Journal of Chemical Physics* **2004**, *120*, 11586.

(223) Møller, C.; Plesset, M. S. *Physical Review* **1934**, *46*, 0618.

(224) Čížek, J. *Journal of Chemical Physics* **1966**, *45*, 4256.

(225) Crawford, T. D.; Schaefer, H. F. *Reviews in Computational Chemistry, Vol 14* **2000**, *14*, 33.

(226) Kállay, M.; Surján, P. R. *Journal of Chemical Physics* **2001**, *115*, 2945.

(227) Kállay, M.; Gauss, J. *Journal of Chemical Physics* **2005**, *123*, 214105.

(228) Bomble, Y. J.; Stanton, J. F.; Kállay, M.; Gauss, J. *Journal of Chemical Physics* **2005**, *123*, 054101.

(229) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *Journal of Chemical Physics* **1991**, *94*, 7221.

(230) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *Journal of Chemical Physics* **1998**, *109*, 7764.

(231) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *Journal of Chemical Physics* **2000**, *112*, 6532.

(232) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *Journal of Chemical Physics* **1999**, *110*, 7650.

(233) Lynch, B. J.; Truhlar, D. G. *Journal of Physical Chemistry A* **2003**, *107*, 3898.

(234) Handy, N. C.; Yamaguchi, Y.; Schaefer, H. F. *Journal of Chemical Physics* **1986**, *84*, 4481.

(235) Gauss, J.; Tajti, A.; Kállay, M.; Stanton, J. F.; Szalay, P. G. *Journal of Chemical Physics* **2006**, *125*, 144111.

(236) Kutzelnigg, W. *Molecular Physics* **1997**, *90*, 909.

(237) Valeev, E. F.; Sherrill, C. D. *Journal of Chemical Physics* **2003**, *118*, 3921.

(238) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *Journal of Chemical Physics* **1992**, *96*, 6796.

(239) Peterson, K. A.; Kendall, R. A.; Dunning, T. H. *Journal of Chemical Physics* **1993**, *99*, 1930.

(240) Moran, D.; Simmonett, A. C.; Leach, F. E.; Allen, W. D.; Schleyer, P. V.; Schaefer, H. F. *Journal of the American Chemical Society* **2006**, *128*, 9342.

(241) Hobza, P.; Šponer, J. *Journal of the American Chemical Society* **2002**, *124*, 11802.

(242) Jurečka, P.; Hobza, P. *Chemical Physics Letters* **2002**, *365*, 89.

(243) Dąbkowska, I.; Jurečka, P.; Hobza, P. *Journal of Chemical Physics* **2005**, *122*, 204322.

(244) Allen, W. D.; East, A. L. L.; Császár, A. G.; In *Structures and Conformations of Non-Rigid Molecules*; Laane, J.; Dakkouri, M., van der Veken, B., Oberhammer, H., Eds.; Kluwer: Dordrecht. **1993**, 343.

(245) Császár, A. G.; Allen, W. D.; Schaefer, H. F. *Journal of Chemical Physics* **1998**, *108*, 9751.

(246) Czakó, G.; Nagy, B.; Tasi, G.; Somogyi, Á.; Šimunek, J.; Noga, J.; Braams, B. J.; Bowman, J. M.; Császár, A. G. *International Journal of Quantum Chemistry* **2009**, *109*, 2393.

(247) Ochterski J. W. *Thermochemistry in Gaussian.* [updated 11 June 2009; cited 2 March 2010]. Available from http://www.gaussian.com/g_whitepap/thermo.htm.

(248) Karton, A.; Martin, J. M. L. *Theoretical Chemistry Accounts* **2006**, *115*, 330.

(249) Klopper, W.; Kutzelnigg, W. *Journal of Molecular Structure (THEOCHEM)* **1986**, *28*, 339.

(250) Blanksby, S. J.; Ramond, T. M.; Davico, G. E.; Nimlos, M. R.; Kato, S.; Bierbaum, V. M.; Lineberger, W. C.; Ellison, G. B.; Okumura, M. *Journal of the American Chemical Society* **2001**, *123*, 9585.

(251) Nagy, P. I.; Dunn, W. J.; Alagona, G.; Ghio, C. *Journal of the American Chemical Society* **1991**, *113*, 6719.

(252) Radom, L.; Lathan, W. A.; Hehre, W. J.; Pople, J. A. *Journal of the American Chemical Society* **1973**, *95*, 693.

(253) Vazquez, S.; Mosquera, R. A.; Rios, M. A.; Van Alsenoy, C. *Journal of Molecular Structure (THEOCHEM)* **1989**, *188*, 95.

(254) Lotrich, V.; Flocke, N.; Ponton, M.; Yau, A. D.; Perera, A.; Deumens, E.; Bartlett, R. J. *Journal of Chemical Physics* **2008**, *128*, 194104.

(255) Janowski, T.; Ford, A. R.; Pulay, P. *Journal of Chemical Theory and Computation* **2007**, *3*, 1368.

(256) Janowski, T.; Pulay, P. *Journal of Chemical Theory and Computation* **2008**, *4*, 1585.

(257) Olson, R. M.; Bentz, J. L.; Kendall, R. A.; Schmidt, M. W.; Gordon, M. S. *Journal of Chemical Theory and Computation* **2007**, *3*, 1312.

(258) Harding, M. E.; Metzroth, T.; Gauss, J.; Auer, A. A. *Journal of Chemical Theory and Computation* **2008**, *4*, 64.

(259) Hughes, T. F.; Flocke, N.; Bartlett, R. J. *Journal of Physical Chemistry A* **2008**, *112*, 5994.

(260) Flocke, N.; Bartlett, R. J. *Journal of Chemical Physics* **2004**, *121*, 10935.

(261) Schütz, M.; Werner, H. J. *Journal of Chemical Physics* **2001**, *114*, 661.

(262) Hampel, C.; Werner, H. J. *Journal of Chemical Physics* **1996**, *104*, 6286.

BIOGRAPHICAL SKETCH

Xiao He was born in Suzhou, Jinagsu Province, China in 1981 and spent childhood and teen ages in Suzhou. He attended Suzhou Middle School for high school education from 1993 to 1999. He obtained his Bachelor of Science degree in physics from the University of Nanjing in 2003 and his Master of Science degree in chemistry from the University of Nanjing in 2006, where he performed research in Professor John Zenghui Zhang's group. In August 2006, he joined the group of Professor Kenneth M. Merz, Jr. for his doctoral studies in physical chemistry at the University of Florida.