IOWA STATE UNIVERSITY
Digital Repository

2009

Slicing, dicing, washing, evolving. All in the name of capturing an electron (or two).

Jonathan Michael Mullin
Iowa State University

**SLICING, DICING, WASHING, EVOLVING. ALL IN THE NAME OF CAPTURING
AN ELECTRON (OR TWO).**

by

Jonathan Michael Mullin

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Physical Chemistry

Program of Study Committee:
Mark Gordon, Major Professor
Amy Andreotti
William Jenks
Robert Jernigan
Xueyu Song

Iowa State University

Ames, Iowa

2009

Dedication

May love and laughter light your days,

and warm your heart and home.

May good and faithful friends be yours,

wherever you may roam.

May peace and plenty bless your world

with joy that long endures.

May all life's passing seasons

bring the best to you and yours!

Table of Contents

Abstract

Chapter one and two introduce the raison d'etre for the research conducted and a background for the methods employed herein. Chapters three and four focus on the study of alanine, and the effects of solvation. Chapters five and six explore fragmentation models from development to general use. Chapter seven focuses on the use of molecular dynamics for solution phase chemical phenomena using the effective fragment potential. Chapter eight provides a general overview and summary of the projects explored in the dissertation.

A BRIEF HISTORY OF SCIENCE AND CHEMISTRY[1]

Chemistry was certainly not the earliest pursuit of modern man. The initial expressions of mankind described little more than experiences of survival. These can be seen in cave paintings and carvings beginning 40,000 years ago, and they have been observed wherever mankind spread.[2] It was not until nearly 20,000 years ago that man made the first known expression of abstract thinking, the Ishango bone. The Ishango bone has a series of hash marks, whose meaning died with its creator. Theories of basic mathematical calculations or a lunar calendar have been proposed, but it is certain that whatever the meaning, the marks represent the beginnings of man's use of abstract symbols to represent ideas.[3]

However, mankind's journey toward scientific discovery took its greatest step forward with the agricultural revolution. Wherever there was plentiful water, grains, and domesticated animals, mankind had the time to think about the world around them.[4] Several examples include India, which by the seventh century had made significant advances, including basic atomic and gravitation theories, along with the development of the Hindu-Arabic numeral system that included zero.[5] Early Chinese cultures created their own numerical systems and the abacus, and documented the first known solar eclipse.[6] In Babylon as early as 1900 BCE, Pythagorean triplets were recorded well before an abstract theory was developed.[7]

Most current scientific traditions, however, can be traced to the Hellenic world in Greece, where Socrates devoted his time to teaching philosophical ideas. It was his students though, especially Aristotle, who took his traditions and pragmatic style to apply

philosophy to the natural world. Aristotle's establishment of the first college is perhaps the contribution most relevant to a dissertation introduction.

Philosophical era

The first chemists, motivated no doubt by practicality, sought an understanding of fire, which held a central part in ancient cultures, as it was used in food and tool preparation. Two ancient Greeks, Democritus and Leucippus, put forth the concept of atoms, atomisim, in 440 BCE, as told by the Roman Lucretius in De Rerum Natura. The co-discovery by the Indian philosopher Kanada also explained an atomicist worldview in the Vaisheshika Sutras. Both theories were purely philosophical in nature and lacked any empirical proof and were discounted in their cultures.[8] Aristotle held the best-known contempt of atomism in 330 BCE.[8b]

Metallurgy and alchemy were also areas of early interest in chemistry. The discovery of alloys is in fact the definition of the Bronze Age. The preciousness of gold gave rise to the study of alchemy in Egypt as early as 2600 BCE.[9] Belief in transmutation was present for some time throughout Europe holding the imagination of philosophers such as Albertus Magnus and Thomas Aquinas in the dark ages of European culture. Muslim scholars such as Avicenna, however, correctly asserted that alchemy was a pseudoscience:

> "Those of the chemical craft know well that no change can be effected
>
> in the different species of substances, though they can produce the
>
> appearance of such change."[9]

It was from the Byzantine capital of Constantinople that science and chemistry continued to be studied during the dark ages in Europe. Due to the work of Muslim

scholars who kept the spirit of scientific inquiry alive, much of what had been lost could be reintroduced in the European renaissance.

Gas Theory / Atomism of Chemistry

During the renaissance, Descartes began conducting experiments and separated science and humanities.[10] Science saw development again in chemistry with Robert Boyle who brought European thinking into the atomism tradition. Boyle was responsible for a theory of gases, which bears his name.[11] Daniel Bernoulli explained Boyle's law of gases via a kinetic theory of gases which rationalized pressure in terms of momentum.[12] Despite Boyle's early work, it is Antoine Lavoisier who is often thought of as the modern father of chemistry. Lavoisier proffered the law of conservation of mass, refuting the phlogiston theory of combustion.[13]

John Dalton's talk in 1803 to the Literary and Philosophical Society in London proposed an atomic view of gases, in which a thin warming shell surrounds spherical atoms. Chemical operations were considered to be the processes of unification or separation of these thin warming shells.

Amadeo Avagadro developed Avagardo's law, which allows atomic weights of gasses to be calculated from their masses.[14] In order to reconcile Dalton's atomic theory with Joseph Louis Gay-Lussac's 1808 law on volumes and the combining of gases, Avagadro explicitly defined molecules and atoms. Initial theories of atomic connections were essentially ionic bonds, though this contradicted Avagadro, who stated that the most basic gases were homogenous diatomic species. In 1860 the first International Chemistry Congress in Karlsruhe, Germany was held. Avagadro's student, Stanislao Cannizzaro, presented a paper on all the atomic weights, which were summarily accepted.[15] This

allowed further classification and led to Dmitri Mendeleev's creation of the periodic table. Of note, Johann Josef Loschmidt first calculated what is currently known as Avagadro's number.[16]

The German chemist August Wilhelm von Hofmann presented to the Royal Institution of Great Britain in 1865 a molecular model, which today would be referred to as ball and stick. Though the geometries were incorrect, his proposed coloring convention is still widely used. In 1874, Jacobus Henricus van 't Hoff and Joseph Achille Le Bel independently proposed the phenomenon of optical activity.[17] Following early work of Jean Biot,[18] they explained the phenomenon by presupposing that chemical bonds between carbon atoms and their neighbors were directed towards the corners of a regular tetrahedron. This led to the development of Emil Fisher's 3-D representation of bonding.[19]

In 1881, in what became a famous Faraday Lecture, the German physicist Hermann Ludwig Ferdinand von Helmholtz came to the conclusion that whatever chemical affinity may be, it would have to be electric in its nature. In the case of salt-like affinities such as KCl or HF, this could be explained. In fact, Svante Arrhenius was the originator of ionic dissociation of a bond in water.[20] Still, nobody had hypothesized how electric forces could contribute to bonding in homogenous diatomics.

In 1898, Ludwig Boltzmann explained the phenomenon of gas phase molecular dissociation, and in doing so drew one of the first rudimentary, yet detailed atomic orbital overlap drawings.[21] Noting first the known fact that molecular iodine vapor dissociates into atoms at higher temperatures, Boltzmann proposed that one must explain the existence of molecules composed of two atoms, whose chemical attraction arose from a

relatively small region on the surface of the atom called the sensitive region, which formed a bond when overlapped.

Electronic Chemistry

Joseph John Thomson discovered, using a Crookes tube, that cathode rays could be deflected by an electric field. He proposed that the rays were not waves, but negatively charge particles he referred to as corpuscles, which later became known as electrons. Thomson observed that corpuscles emerged from the atoms in an electrode, and thus atoms were divisible. His representation of the atom is referred to as the plum pudding model, in which the corpuscles were distributed in a uniform cloud of positive charge.[22]

In Ernest Rutherford's gold foil experiment, a small portion of the alpha particles were deflected, indicating a small, concentrated positive charge, giving rise to a model of the atom resembling the planetary orbits.[23] A main point of both of these experiments is that chemical activity comes from the electrons and not the nucleus. Electronic motion gives rise to observable chemical properties such as emission spectra, characterized by Gustav Kirchhoff and Robert Bunsen. Rutherford's model was able to reproduce Johann Jakob Balmer's hydrogen emission spectrum.[24]

In 1916 Niels Bohr explained the orientation of the electrons around the nucleus mathematically via an exclusion principle.[25] Wolfgang Pauli was able to expand on his idea by defining an unknown state, electron spin, which allowed *a priori* orbital definitions along with periodicity of the elements.[26] Concurrently, Arnold Sommerfeld had created a classical quantum theory, which was able to predict some chemical properties.[27] However, the phenomenon that holds atoms together was still unexplained. Ideas of valence attempted to explain atomic bonds. Richard Abegg first postulated a rule

of eight, leading to the idea of valence shells.[28] Gilbert N. Lewis developed the idea of electron pairs as the basis of chemical bonds, which is described in his 1923 book *Valence and the Structure of the Atom*:[29]

> ''The new theory, which includes the possibility of complete ionization as a special case, may be given definite expression as follows: Two atoms may conform to the rule of eight, or the octet rule, not only by the transfer of electrons from one atom to another, but also by sharing one or more pairs of electrons. The electrons which are held in common by two atoms may be considered to belong to the outer shell of both atoms.''

Quantum Chemistry

The age of quantum chemistry began in 1926 when Erwin Schrödinger took the basic wave equation and enforced boundary conditions to reproduce the hydrogen spectra.[30] Walter Heitler and Fritz London published the first calculations on H_2 in 1927.[31] This first phase of quantum descriptions of chemical bonding lasted from 1927-1935. In this era several attempts were made to rectify known valence theory with quantum calculations. Linus Pauling's book, *The Nature of the Chemical Bond*,[32] suggested the concept of hybridization and Valence Bond (VB) theory. Concurrently, Robert Mulliken and Friedrich Hund attempted to understand diatomic molecular spectra. Their explanation of molecular orbitals (MOs) allowed for correlation diagrams and qualitative understanding of MO energies.[33] Extension of this method by John Lennard-Jones represented bonding MOs as a linear combination of atomic orbitals (AOs).[34] With

the development of quantum theory, the nature of the chemical bond could finally be explained, though it took nearly thirty years for a generally accepted answer.

John Slater used the virial theorem,[35] in order to show that as two atoms come together the potential energy will decrease and the kinetic energy will increase; therefore, the electrons will accumulate in the bonds because of the overlap of orbitals, suggesting that a lowering of the potential energy is the cause of the bonding.[36] Hans G. A. Hellman noticed that in the Heitler-London formulation, the potential energy increased and the kinetic energy decreased.[37] He was the first to realize that the basis of chemical bonding came from the lowering of the kinetic energy. Klaus Ruedenburg showed in his 1962 paper and subsequent papers[38] that covalent bonding arises mainly from the constructive interference contribution to the kinetic energy. In a bonding system there is localization near the nuclei; this lowers the potential energy and raises the kinetic energy. Delocalization into the bond region increases the potential energy and lowers the kinetic energy. This lowering of the kinetic energy is the origin of the chemical bond. The lowering of the potential energy, via the wavefunction contraction, is not the ultimate source of chemical bonding, as previously believed. Theoretical chemistry has now moved on to describe the basis of several conceptual theories such as: hydrogen bonding, charge transfer, bond polarity, ring strain and other chemical phenomena which are still the pursuit of modern theoretical chemists.

References:

1. Ruedenburg, K. discussions and notes on his presentations.
2. Marshack, A. The Roots of Civilization, Colonial Hill, Mount Kisco, NY, **1991**.
3. (a) Darling, D.J. The Universal Book of Mathematics: From Abracadabra to Zeno's Paradoxes. John Wiley and Sons, **2004**. (b) Williams, S. W.: "Mathematicians of the African Diaspora" http://www.math.buffalo.edu/mad/Ancient-Africa/ishango.html. Accessed

Sept. 2008

4. Diamond, J. Guns, Germs, and Steel: The Fates of Human Societies. W.W.Norton, **1997**.

5. Flegg, Graham. *Numbers: Their History and Meaning*. Courier Dover Publications, **2002**.

6. Peng, Y. H. *Li, Qi and Shu: An Introduction to Science and Civilization in China*, Courier Dover Publications, **2000**.

7. Hoffman, P. *The man who loved only numbers: the story of Paul Erdös and the search for mathematical truth*, (New York: Hyperion), **1998**, p.187.

8. (a) Lloyd, G. *Early Greek Science: Thales to Aristotle*. London; New York: Chatto and Windus; W. W. Norton & Company. **1970**, 108–109. (b) Teresi, Dick *Lost Discoveries: The Ancient Roots of Modern Science*. Simon & Schuster. **2003**. pp. 213–214.

9. Neumann, E. *The origins and history of consciousness*, with a foreword by C.G. Jung. Translated from the German by R.F.C. Hull. New York: Pantheon Books, **1954**.

10. Descartes, R. Laurence J. Lafleur (trans.), *Discourse on Method and Meditations*. New York: The Liberal Arts Press. **1960**.

11. Boyle, R., *The Sceptical Chymist*, **1661**.

12. Bernoulli, D., *Hydrodynamica*, **1738**.

13. Lavoisier, A., *Elements of Chemistry*(1738). New York: Dover.- Reprint of Robert Kerr's English translation of **1790**.

14. Avagadro*, A. Essay on Determining the Relative Masses of the Elementary Molecules of Bodies*, **1811**.

15. Cannizzaro, S. *Sunto di un corso di Filosofia chimica*, **1858**.

16. Loschmidt, J. J. *Chemische Studien I* **1861**.

17. van 't Hoff, J. H. *a chimie dans l'éspace* **1874**.

18. Biot, J. *Traité de physique expérimentale et mathématique*, Deterville, **1816**.

19. Kunz, H. *Angewandte Chemie International Edition* **2002**, *41*, 4439.

20. Arrhenius, S. *Recherches sur la conductivité galvanique des électrolytes*, doctoral dissertation, Stockholm, Royal publishing house, P.A. Norstedt & söner, **1884**.

21. Boltzmann, L. *Lectures on Gas Theory*, 1898, Dover **1985**.

22. Thomson, J. J. *Philosophical Magazine, **1904**, series 7(39),* 237.

23. Geiger H.; Marsden E. *Proc. Royal Soc., **1909**, A82*, 495.

24. Rutherford E., *Philosophical Magazine*, **1911**, *series 6(21),* 669.

25. (a) Bohr, N. *Philosophical Magazine*, **1913**, *26*, 1. (b) Bohr, N. *Philosophical Magazine*, **1913**, *26*, 476. (c) Bohr, N. *Philosophical Magazine*, **1913**, *26*, 857. (d) Bohr, N. *Nature*, **1914**, *92*, 231. (e) Bohr, N. *Nature*, **1921** *106*, 104.

26. Wolfgang,P. *Exclusion principle and quantum mechanics; lecture given in Stockholm after the award of the Nobel prize of physics,*1945.Published: Neuchcatel, Editions du Griffon, **1947**.

27. Sommerfeld, A. *Atombau und Spektrallinien,* Friedrich Vieweg und Sohn, Braunschweig, **1919**.

28. Richard, A. *J. Am. Chem.* **1910**, *43*, 563.

29. Lewis, G. N*, Valence and the Structure of the Atom*, **1923**.

30. (a) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 361. **(**b) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 489. **(**c) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 734. **(**d) Schrödinger, E. *Ann. Physik.* **1926**, 80, 437. **(**e) Schrödinger, E. *Ann. Physik.* **1926**, *81*, 109.

31. Heitler, W.; London,F. *Z. Physik*, **1927**, *44*, 455.

32. Pauling, L. *The Nature of the Chemical Bond*, Cornell University Press, 1960.

33. Soukup, R. W. *Monatsh Chem*, **2005**, *136*, 803.

34. Lennard-Jones, J. E. *Trans.Faraday Soc.* **1929**, *25*, 668.

35. Slater, J.C. *Phys. Rev.* **1931**, *37*, 481.

36. Slater, J. C. *J. Chem. Phys*. **1933**, *1*, 687.

37. Kutzelnigg, W. *Angew Chem*, **1973**, *13*, 556. (b) Hellmann, H.; Jost, W. *Z. Elektrochem Angew. Phys. Chem.*, **1934**, *40*, 806. (c) Jost, W., *Z. Elektrochem. Angew. Phys. Chem.* **1935**, *41*,66.

38. (a) Ruedenberg, K. *Rev. Mod. Phys.* **1962**, *34*, 326. (b) Edmiston, C.; Ruedenberg, K. *Rev Mod Phys* **1963**, *35*, 457. (c) Edmiston, C.; Ruedenberg, K. *J Chem Phys*, **1965**, *43*, 97. (d) Thompson, H.B.; *Inorg Chem* **1968**, *7*, 604. (e) Diner, S.; Malrieu, J. P.; Claverie, P. *Theor Chim Acta*, **1969**, *13*, 1. (f) England, W.; Salmon, L.S.; Ruedenberg, K. *Top Current Chem*, **1971**, *23*, 31. (g) Gordon, M.; Jensen, J. *Theor Chem Acc,* **2000**, *103*, 248.

GENERAL INTRODUCTION

I. General Overview

Solution phase chemistry is a major component of many life forms on earth. When considering biological systems the interactions of water with the solute are involved many reactions. Initial studies contained herein consider how the effects of water on the solute change chemical properties of the solute. Both discrete and continuum models of solvation are explored to help quantify when each are useful. The discrete method employs Monte Carlo simulations to sample the solvent configuration space, and molecular dynamics simulations to study bulk properties. For alanine, the exploration of discrete solvation shows the transition from neutral to zwitterion form and finally to a full solvation shell. Further development work on molecular dynamics shows the promise of the effective fragment potential as a means to study the properties of solvents other than water. The chapters on the systematic fragmentation method show how the correct physical treatment of the chemical subunits and the interactions between them allow for the fragmentation of a molecule in a way that gives nearly linear scaling with an accuracy of ~1 kcal/mol.

II. Theoretical Methods

This section presents a brief overview of the theoretical methods that will be considered throughout later chapters. Within this dissertation three main types of models are used; quantum mechanics, molecular mechanics and fragmentation models. Quantum mechanics (QM) describes a molecule in terms of the probability density of the electrons surrounding the positively charged nuclei. Molecular mechanics (MM) treats a molecule as a collection of atoms held together by (often rigid) bonds with the energies calculated via

classical potential energy functions. Fragmentation models allow for large systems to be broken into smaller, chemically meaningful, fragments, treated individually with QM, and then added back together in order to capture the physical properties of the composite system. Much of the work described herein combines these methods to solve problems that are inaccessible by a fully QM approach. A brief description of each method will be given below.

A. Quantum Mechanics (QM) / Schrödinger Equation

Since electrons display both wave and particle characteristics, they must be described in terms of a wavefunction, Ψ. The time-dependent Schrödinger[3] equation postulates that the state of Ψ will change with time as:

$$\mathrm{H}(r,t)\Psi(r,t) = i\hbar\frac{\partial\Psi(r,t)}{\partial t} \tag{1}$$

where $\mathrm{H}(r,t)$ is the Hamiltonian operator, i is $(-1)^{1/2}$, and $\hbar = \dfrac{h}{2\pi}$

A simplification of this procedure is available when the time variable is separable from the spatial variables. This may be expressed for our purposes as the potential energy of the system being time independent. This simplification, that is used in most electronic structure methods, is the time independent Schrödinger Equation.[1]

$$H\Psi = E\Psi \tag{2}$$

Eq. (2) is an eigenvalue equation. The Hamiltonian operator is composed of both kinetic (T) and potential (V) energy operators for all particles:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{Ne} + \hat{V}_{ee} \tag{3}$$

where $\hat{T}_N$ is the operator for the nuclear kinetic energy (KE), $\hat{T}_e$ is the operator for the electronic KE, $\hat{V}_{NN}$ represents the nuclear-nuclear repulsive potential energy (PE), $\hat{V}_{Ne}$ is the nuclear-electronic attraction PE, and $\hat{V}_{ee}$ is the electron-electron repulsion PE.

A common approximation in quantum mechanics is the Born-Oppenheimer[2] approximation. Since a single proton is ~1800 times heavier than electrons, it is a good approximation that electronic motion will instantaneously follow the nuclear motion. This allows an electronic-nuclear separation of the Hamiltonian. Within the Born-Oppenheimer approximation, the nuclei are assumed to remain fixed when one solves for the electronic motion (thus, $\hat{T}_N = 0$ and $\hat{V}_{NN}$ is a constant). This allows the motions of the nuclei and electrons to be treated separately, due to the order of magnitude difference in their respective masses. The Schrödinger equation then may be written as:

$$H_e(R)\Psi_e(R) = E_e(R)\Psi_e(R) \tag{4}$$

where R represents the nuclear coordinates and $H_e(R)$ is the electronic Hamiltonian given (in atomic units) by:

$$H_e = -\frac{1}{2}\sum_i^n \nabla_i^2 - \sum_i^n \sum_A^N \frac{Z_A}{r_{iA}} + \sum_i^n \sum_{j<i}^n \frac{1}{r_{ij}} \tag{5}$$

where n is the number of electrons, N is the number of nuclei, ∇_i^2 is the Laplacian operation for the ith electron, r_{iA} and r_{ij} are the distances between electron i and nucleus A, and between the ith and jth electron respectively. Z_A is the nuclear charge of the nucleus A.

A further approximation is needed to deal with separation of coordinates in the r_{ij} or electron-electron repulsion term in Eq. (5). This is done by using a mean field approximation in Hartree Fock theory,[3] in which each electron interacts with an average field representing all other (n-1) electrons.

B. HF (Hartree Fock)

Hartree Fock theory has become the base level of most electronic structure calculations. This is due to its qualitative reliability and computational feasibility. HF is based on the independent electron approximation, in which an antisymmetrized product of one-electron functions (spin orbitals ψ) gives the electronic wave function. Due to the antisymmetry requirement (Pauli Principle) for electrons, the wavefunction is expressed as a Slater determinant:

$$\Psi_e = (n!)^{-\frac{1}{2}} \begin{vmatrix} \psi_1(e_1) & \psi_2(e_1) & \dots & \psi_i(e_1) \\ \psi_1(e_2) & \psi_2(e_2) & \dots & \psi_i(e_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(e_i) & \psi_2(e_i) & \dots & \psi_i(e_i) \end{vmatrix} \tag{6}$$

In Eq.(6) ψ_i is the ith molecular spin orbital and e_i is the ith electron. The variational principle when applied to Eq. (6) leads to the Hartree-Fock equations:

$$\hat{F}\psi_i = \varepsilon_i \psi_i \tag{7}$$

where $\hat{F}$ is the Fock operator , ε_i is the energy of the ith orbital ψ_i.

The Fock operator depends on the orbital ψ_i and therefore Eq. (7) must be solved iteratively to self-consistency. An initial guess of ψ_i facilitates the process. This initial guess is often obtained using Huckel theory.[4] $\hat{F}$ is thus constructed and Eq. (7) is used to define a new set

of ψ_i, which begets a new $\hat{F}$. This cycle is repeated until the orbitals used to construct $\hat{F}$ are the same as those obtained from solving Eq. (7).

As analytic solutions are preferred to numerical solutions, it is the expression of ψ_i as a linear combination of atomic orbitals (LCAO)[5] that facilities an analytic solution.

$$\psi_i = \sum_{\mu} C_{\mu i} \chi_{\mu} \tag{8}$$

where the $C_{\mu i}$ are LCAO coefficients, and the χ_{μ} are atomic orbitals (basis functions). An infinite basis set limit gives the exact HF wavefunction and energy. However, since one typically employs a finite basis set, the LCAO approach is an approximation. These atomic orbitals χ_{μ} are usually expressed as a linear combination of Gaussian functions:

$$\chi_{\mu}(r) = \sum_{p} N_p x_p^l y_p^m z_p^n e^{-\alpha_p r^2} \tag{9}$$

where N_p is a normalization constant l, m, n are integers whose sum is the angular momentum, and α_p is a coefficient which determines the size of the Gaussian function.

Correlation of electrons does exist in the Hartree Fock method for electrons of the same spin. This correlation is due to the Fermi hole, or Pauli exclusion principle, which requires electrons of like spin to remain apart; however, electrons of opposite spin are not correlated. Electron repulsion in HF is overestimated due to this lack of electron correlation. The lack of electron-electron correlation of opposite spin electrons results in an energy difference between the energy of the HF limit and the exact non-relativistic energy. This difference is often referred to as the electron correlation energy.

There are many approaches to add electron correlation effects to the HF wavefunction. These include configuration interaction (CI), perturbation theory (PT),

coupled cluster (CC) theory, and for short-range correlation effects density functional theory (DFT). Those methods, which have been utilized in the dissertation, will be discussed below.

C.DFT (Density Functional Theory)

Density functional theory[6] (DFT) is based upon the fact that the ground state electronic energy of a system can be described via the electron density. This is known as the Hohenberg-Kohn[7] theorem. It is attractive because the density p depends on only three coordinates, whereas the wavefunction depends on 3n coordinates for n electrons. The transformation of the electron density to any property including the electronic energy is unique, yet the exact form of such a density functional remains unknown. The holy grail of DFT is to design a universal functional E[p] that expresses the energy as a function of the exact density. The most common approximation is based on the Kohn-Sham approach (similar to the independent particle model in Hartree-Fock theory) in which one constructs exchange-correlation functionals that are designed to mimic the exact exchange and correlation energies. There are five hierarchal rungs defined as the DFT "Jacob's ladder" by Perdew in which functionals may be classified.[8] The first rung is the local density approximation (LDA), in which the energy is expressed only in terms of the electron density. The second rung is the generalized gradient approximation (GGA), in which the energy depends on both the electron density and the density gradient. The third rung is the meta-GGA, in which some parameterized fraction of HF exchange is taken to be a component of the exchange functional. The fourth rung is the hyper GGA, adds the exact exchange energy density, a nonlocal functional of occupied Kohn-Sham orbitals. This extends the GGA approximation in such a way that the energy depends on the electron density, the density gradient, and the kinetic energy density. Hybrid functions are in a sense fourth rung, though

they do not use the exact exchange. The fifth rung extends the hyper GGA by combining exact exchange with exact partial correlation, thereby depending on both the occupied and unoccupied Kohn-Sham orbitals. This dissertation employs the fourth rung via B3LYP[9] chosen due to the popularity of the B3LYP functional among chemists.

D. MPn (Perturbation Theory)

Perturbation theory[10,11] improves the energy relative to that of a reference wave function (e.g., HF) by including dynamic electron correlation via a perturbing Hamiltonian, H_{PT}, which should be small perturbation of the zeroth order Hamiltonian H_0. This may be described mathematically by defining the perturbing Hamiltonian in two parts: the larger unperturbed component from HF theory, and the smaller (often external) perturbation, V, where λ is an arbitrary real parameter.

$$H_{PT} = H_0 + \lambda V \qquad (10)$$

Here we consider Rayleigh-Schrödinger perturbation theory through second order. A Taylor series is employed to expand the energy and wave function using the n-th state HF wavefunction, $\Psi_i^{(0)}$.

$$\begin{aligned} \Psi_i &= \Psi_i^{(0)} + \lambda \Psi_i^{(1)} + \lambda^2 \Psi_i^{(2)} + ... \\ E_i &= E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + ... \end{aligned} \qquad (11)$$

Collating powers of λ and substituting equation (11) into the Schrödinger equation Eq. (2) yields:

$$\begin{aligned} H_0 |i\rangle &= E_i^{(0)} |i\rangle \\ H_0 |\Psi_i^{(1)}\rangle + V|i\rangle &= E_i^{(0)} |\Psi_i^{(1)}\rangle + E_i^{(1)} |i\rangle \\ H_0 |\Psi_i^{(2)}\rangle + V|\Psi_i^{(1)}\rangle &= E_i^{(0)} |\Psi_i^{(2)}\rangle + E_i^{(1)} |\Psi_i^{(1)}\rangle + E_i^{(2)} |i\rangle \end{aligned} \qquad (12)$$

Multiplying these equations by the zeroth order waver function and integrating over all space yields the energy in terms of the perturbation.

$$
\begin{aligned}
E_i^{(0)} &= \langle i | H_0 | i \rangle \\
E_i^{(1)} &= \langle i | V | i \rangle \\
E_i^{(2)} &= \langle i | V | \Psi_i^{(1)} \rangle
\end{aligned}
\tag{13}
$$

This expansion is then truncated at the nth order give the nth order perturbation.

$$
E^{MPn} = E_0^{(0)} + E_0^{(1)} + E_0^{(2)} \cdots + E_0^{(n)}
\tag{14}
$$

In second order Møller-Plesset[11] (MP) perturbation theory the perturbation takes the form:

$$
V = H - F - \langle \Phi_0 | H - F | \Phi_0 \rangle
\tag{15}
$$

where Φ_0 is a normalized Slater determinate of the Fock operators lowest eigenfunction. The unperturbed Hamiltonian takes the form of the shifted Fock operator:

$$
H_o = F + \langle \Phi_0 | H - F | \Phi_0 \rangle .
\tag{16}
$$

Since the Slater determinate Φ_0 is an eigenfunction of the Fock operator (F), the zeroth-order energy is the HF energy:

$$
E_{MP0} = E_{HF} = \langle \Phi_0 | H | \Phi_0 \rangle .
\tag{17}
$$

The first order MP energy is formally zero.

$$
E_{MP1} = \langle \Phi_0 | V | \Phi_0 \rangle
\tag{18}
$$

Therefore MP correlation comes from the second order energy. The MP2 formulation is based on the double excitation from occupied to virtual orbitals.

$$
E_{MP2} = -\frac{1}{4} \sum_{ij} \sum_{ab} \frac{|\langle ia \| jb \rangle|^2}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j}
\tag{19}
$$

Where:

$$\left|\langle ia\|jb\rangle\right| = \langle ia|jb\rangle - \langle ib|ja\rangle, \text{ and } \langle ia|jb\rangle = \int dx_1\, dx_2\, x_i^*(x_1) x_a^*(x_2) r_{12}^{-1} x_j(x_1) x_b(x_2). \quad (20)$$

Where a,b are virtual orbitals and i,j are occupied orbitals. The work discussed in this thesis will employ second order perturbation theory (MP2) in order to recover most of the dynamic correlation energy contribution.

E. CCSD(T) (Coupled Cluster Singles, Doubles, and Perturbative Triples)

Coupled cluster (CC) theory is a many body approach to obtaining the electron correlation correction to HF theory. In perturbation theory all types of corrections (one particle, two particle, etc.) are added to a given order, to the reference wavefunction.[12] Conversely, CC methods attempt to include all corrections of a given type to an infinite order. The coupled cluster wavefunction Ψ_{CC} can be written as

$$\Psi_{CC} = e^T \Psi_{HF} \quad (21)$$

The coupled cluster operator T is given by:

$$T = T_1 + T_2 + T_3 \cdots + T_n. \quad (22)$$

Where T_i may be defined via second quantization:

$$T_1 = \sum_i \sum_a t_i^a \hat{a}_i \hat{a}_a^\dagger$$

$$T_2 = \frac{1}{4} \sum_{i,j} \sum_{a,b} t_{ij}^{ab} \hat{a}_i \hat{a}_j \hat{a}_a^\dagger \hat{a}_b^\dagger \quad (23)$$

$$T_3 = \frac{1}{8} \sum_{i,j,k} \sum_{a,b,c} t_{ijk}^{abc} \hat{a}_i \hat{a}_j \hat{a}_k \hat{a}_a^\dagger \hat{a}_b^\dagger \hat{a}_c^\dagger$$

where i,j,k are occupied orbitals and a,b,c are unoccupied orbitals. The $\hat{a}$ and $\hat{a}^\dagger$ are creation and annihilation operators and t_i^a are the unknown coefficients for Ψ_{CC}. The T_i

operator acts on the HF wavefunction Ψ_{HF} to give all of the i-particle interactions. Therefore, If Eqn. 14 is truncated at the triples term T_3, the method is CCSDT, adding a quadruple yields CCSDTQ and so on. There are several hybrid approaches to truncate at a lower order to preserve accuracy and computational efficiency. One of the most popular is CCSD(T) in which the triples contribution is evaluated by perturbation theory and then added to the CCSD result.

F. Molecular Mechanics

In contrast to QM methods, molecular mechanics (MM) does not explicitly consider electrons. Rather, molecules are considered to be a collection of points (atoms) connected by a potential represented by a spring. Thus, no Schrödinger equation needs to be solved, and most MM methods scale linearly to quadratically. Dynamical properties are treated via Newtonian mechanics. The energy is expressed as a sum of stretching, bending, and torsional motions, plus non-bonded atom-atom (Van der Waals and electrostatic) interactions, along with the corresponding cross terms. Parameters (e.g., vibrational force constants) for these terms are set to experimental data or to more accurate (e.g., MP2) computational data

Non-bonded interactions, such as those found in liquids, solvated molecules or between different parts of a large molecule, are a focus of this dissertation. Non-bonded interactions can be composed of Van der Waals interactions, electrostatic interactions, and exchange repulsion. The Van der Waals energy describes the attraction between atoms that are not covalently bonded. The Lennard-Jones (LJ) potential[13] incorporates an approximation to the Van der Waals energy in an r^{-6} term.

$$V^{LJ}(r) = 4e\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{24}$$

where σ and ε are LJ parameters and r describes the inter-atomic distance. This potential gives the strong, repulsive short-range interaction (r^{-12}), as well as the long-range attraction (r^{-6}). The latter is the lead term in the dispersion expansion (See Eq. 29).

The internal distribution of negatively charged electrons and positively charged nuclei gives rise to electrostatic interactions. The simplest approach to describing such interactions is to assign a point charge to each point (typically an atom) in a MM model. The Coulomb potential depends on the interaction between point charges:

$$V^{el}(r_{ab}) = \frac{q_a q_b}{r_{ab}} \tag{25}$$

where q_a and q_b are the charges on points a and b, respectively, and r_{ab} is the distance between those points.

Many common solvent methods (for water) describe the interaction potential as a sum of Coulomb interactions between all charges and a single Lennard-Jones interaction between oxygen atoms.

$$V = \sum_{ab} \frac{q_a q_b}{r_{ab}} + 4\varepsilon_0\left[\left(\frac{\sigma_o}{r_{00}}\right)^{12} - \left(\frac{\sigma_o}{r_{00}}\right)^{6}\right] \tag{26}$$

In Eq. (26), q_a and q_b are again the charges on each atom a and b, in the system, r_{ab} is the distance between points a and b, r_{OO} is the distance between two oxygen atoms, and ε_O and σ_O are parameterized for each model.

Polarization has been included in several modern force fields, where an induced dipole moment (P) of a point is related to the electric field(E) via the electronic polarizability(α).

$$P = \alpha E$$

Implementations have been based on point charges,[14] density,[15] distributed multipoles,[16] among other approaches.[17]

G. EFP (Effective Fragment Potential)

The effective fragment potential (EFP) model has been implemented at the HF,[18] DFT,[19] and MP2[20] levels of theory. This has been accomplished using the EFP1 formalism, in which the exchange repulsion and dispersion terms contain fitted parameters. The EFP1 method has been developed only for water. A general EFP2[21] implementation contains no fitted parameters and is therefore applicable to any species. The EFP1 method contains one-electron potentials, which are added to the *ab initio* electronic Hamiltonian of the solute. The first of these terms represents the Columbic (electrostatic) interactions between two fragments (EFP-EFP) or between a fragment and quantum mechanical molecule (EFP-QM), screened by a charge penetration function that corrects for overlapping electron densities. The second term represents the induction (polarization) EFP-EFP or EFP-QM interaction. The third term is fitted to the water dimer potential. It contains those terms not accounted for in the first two: exchange repulsion, charge transfer. Dispersion is treated as a separate fourth term in the MP2 version. The first two terms are determined directly from QM calculations on the monomer. The EFP1 formulation for an EFP solvent molecule μ and a QM coordinate s as follows:

$$V_{el}(\mu,s) = \sum_{k=1}^{K} V_k^{Elec}(\mu,s) + \sum_{l}^{L} V_l^{Pol}(\mu,s) + \sum_{m}^{M} V_m^{\mathrm{Re}\,m}(\mu,s)$$

(27)

The three terms on the right hand side of Eq. (27) represent the Coulomb, induction, and remainder terms respectively, where K, L and M are the number of corresponding expansion points.

The Coulomb interaction is represented by a distributed multipolar analysis[22] (DMA) of the molecular density, using multipoles through octopole moments at K=5 expansion points for the water molecule (nuclear centers and bond midpoints). The DMA is a pointwise model, thus it cannot account for the overlap of the charge densities between two molecules, as they approach each other. Correcting for this quantum effect requires the coulomb potential to be multiplied by a distance dependent cutoff function (charge penetration term), which is added to the EFP-EFP charge-charge interaction,[23] and to the EFP-QM interaction.

The polarization/induction term in Eq. (26) is treated by a self-consistent scheme using distributed localized molecular orbital (LMO) polarizabilities. The molecular polarizability tensor is expressed as a tensor sum of the LMO polarizabilities, centered at the LMO centroids. For water, four such LMOs are used: two O lone pairs, and two O-H bonds. The polarization energy is then iterated to self-consistency within the SCF cycles.

The third term in Eq. (26) consists of exchange repulsion, charge transfer, and some short-range correlation contribution (in the DFT version). The dispersion (in the MP2 version), is a separate term. This term for the EFP-QM region is represented as a linear combination of Gaussian functions expanded at the atom centers. For the EFP-EFP interaction a single exponential is used and the expansion is done at the atom centers and the center of mass, in order to better capture the angular dependence of the charge transfer

contribution. The coefficients and exponents of the Gaussian and exponential functions were optimized, by fitting to water dimer structures, chosen to represent a selection of water-water orientations and O-O distances.

The generalized effective fragment potential (EFP2) method can be used to represent any molecule of interest. In EFP2 all terms are derived from first principles with no empirically fitted parameters. The interaction energy includes Coulomb, polarization, exchange-repulsion,[18] dispersion,[19] and charge transfer.[24]

$$E = E_{coul} + E_{ind} + E_{exrep} + E_{disp} + E_{ct}. \tag{28}$$

The exchange repulsion interaction between two fragments is derived as an expansion in the intermolecular overlap. When this overlap expansion is expressed in terms of frozen LMOs on each fragment, the expansion can reliably be truncated at the quadratic term. This term does require each EFP to carry a basis set. Since the same basis set is used to generate the multipoles and the molecular polarizability tensor, EFP calculations are basis set dependent. The dependence of the computational cost of an EFP calculation on the basis set appears primarily in the initial generation of the EFP. Since the basis set is used only to calculate overlap integrals, the computation is very fast and quite large basis sets are realistic.

Dispersion interactions are often expressed by an inverse R expansion,

$$E_{disp} = \sum_n C_n R^{-n} \tag{29}$$

where the coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities integrated over the entire frequency range.[19] The first term in the expansion, n=6, corresponds to the induced dipole-induced dipole (Van der Waals) interactions. In the

EFP2 method, this term is evaluated using the time-dependent HF method. In addition the contribution of the n=8 term is estimated. The C_6 coefficients are derived in terms of interactions between pairs of LMOs on the two interacting fragments.

The charge transfer interaction is derived using a supermolecule approach, in which the occupied valence molecular orbitals on one fragment are allowed to interact with the virtual orbitals on another fragment. This interaction term leads to significant energy lowering in *ab initio* calculations on ionic or highly polar species when incomplete basis sets are employed. An approximate formula[24] for the charge transfer interaction in the EFP2 method was derived and implemented using a second order perturbative treatment of the intermolecular interactions for a pair of molecules at the Hartree–Fock level of theory. To date, the EFP-QM interactions have not been implemented for dispersion or charge transfer.

H. PCM (polarizable continuum model)

For theoretical studies of condensed phase chemistry, continuum solvent models have been employed in various forms.[26] Continuum models treat the bulk solvent as a dielectric medium without distinct structure. This is different from discrete models like EFPs in which individual molecules are physically represented. Instead, in continuum models a distribution of charges is used represent the bulk solvent. The polarizable continuum model (PCM)[27,28] calculates this charge distribution using *ab initio* electronic structure theory. The solvent-solute interface is described by a set of interlocking spheres, which are centered, on individual atoms. These spheres are constructed from arbitrarily small interlocking geometric shapes (tesserae). Poisson statistics are then employed for each tessera, thus yielding an apparent surface charge (ASC) at the center of each tessera. Through parameterization of the

solute cavity, this model can generally reproduce experimental solvation energies to within a few kcal/mol.[27,28]

In order to actually calculate the ASC a vector, q, is determined by the matrix equation:

$$Cq = g \tag{30}$$

where the vector g is a function of the solute electrostatic potential vector, V, and C is a geometric matrix. Both g and C have different forms for different tessellation methods and different PCM formalisms. The implementation in GAMESS is based on GEPOL,[29] which uses triangular tesserae. For the isotropic integral equation formalism PCM (IEF-PCM)[26] the equations describing C and g are:

$$C = \left(\frac{A}{2} - D \right)^{-1} \left(\frac{\varepsilon+1}{\varepsilon-1} \frac{A}{2} - D \right) A^{-1} S A^{-1} \tag{31}$$

$$g = -V \tag{32}$$

Where:

$$A_{ii} = a, \qquad A_{ij} = 0 \tag{33}$$

$$S_{ii} = 1.07 a_i \sqrt{4\pi a_i}, \qquad S_{ij} = \frac{a_i a_j}{|\vec{r}_i - \vec{r}_j|} \tag{34}$$

$$D_{ii} = -\frac{1.07 a_i \sqrt{4\pi a_i}}{8\pi R_i}, \qquad D_{ij} = \frac{a_i a_j (\vec{r}_i - \vec{r}_j) \hat{n}_j}{4\pi |\vec{r}_i - \vec{r}_j|^3} \tag{35}$$

and $\vec{r}_i$, a_i , $\hat{n}_j$, R_i and V are, respectively, the center, area, orthogonal unitary vector, the sphere radius an solute electrostatic potential vector for tessera i .

I. MC (Monte Carlo)

The Monte Carlo methodology employed uses a Metropolis sampling method.[30] The Monte Carlo simulation causes one or more of the coordinates to be displaced in a random manner, after which the energy of the displaced system is evaluated. Acceptance of this geometry is guaranteed if the energy is lower than the previous geometry, while those geometries with a higher energy are accepted with a probability determined by the Boltzmann factor:

$$\text{Probability} = e^{(-E(r_i)/k_B T)} \tag{36}$$

where k_B is the Boltzmann constant and T is the temperature.

This procedure tends to select for energies that are only slightly higher than those in the previous steps. Simulated annealing.[31] allows for the temperature to decrease systematically from a given initial value. A high initial temperature allows for greater freedom in exploring the potential energy surface, while the successive steps down in temperature should find the low energy structures. A local minimization procedure, basin hopping,[31] may be used intermittently, allowing the Monte Carlo simulation to jump from local minimum to local minimum.

J. MD (Molecular Dynamics)

Molecular dynamics simulations solve the Newtonian or Lagrangian equations of motion, in order to predict bulk properties of molecular systems. A system comprised of QM or MM or both is solved for a given time length. Early implementations used LJ potentials described in Eq. (24). Within the GAMESS program a MD code has been implemented for use with the

EFP and fist principles methods (HF, DFT, MP2). Integration (Figure 1) in the time domain is done via either the leapfrog[32] or velocity Verlet[33] method.

The original Verlet method[34] is a direct implementation of Newton's second-order equations of motion. Considering the positions($r(t)$) and acceleration($a(t)$), along with the positions of the previous step($r(t-\delta t)$), the positions at $r(t+\delta t)$ can be elucidated. The Verlet equation for advancing in time is given by,

$$r(t+\delta t)=2\,r(t)-\,r(t-\delta t) + \delta t^2 \bullet a(t). \tag{37}$$

The Verlet methods may introduce numerical noise, because they add the small acceleration term to the difference between two large terms that depend on position, in order to generate a trajectory.[35] Velocities are not directly considered in the Verlet approach. The leapfrog integration uses a half time-step in order to calculate positions and velocities as follows:

$$r(t+\delta t)= r(t) + \delta t \bullet v(t+1/2\delta t),$$

$$v(t+\delta t)= v(t-1/2\delta t) + \delta t \bullet a(t). \tag{38}$$

The velocity Verlet method is able to minimize rounding error by storing the positions, velocity and acceleration at the same time, t. The velocity Verlet equation has the form:

$$r(t+\delta t)= r(t) + \delta t \bullet v(t) + 1/2\ \delta t^2 a(t),$$

$$v(t+\delta t)= v(t) + 1/2\delta t[\ a(t) + a(t+\delta t)]. \tag{39}$$

Both the leapfrog and velocity Verlet methods are algebraically equivalent to the original Verlet method.

Currently, three ensembles have been implemented in the GAMESS MD code: (1) The microcanonical ensemble (NVE) has a constant number of molecules (N), volume (V) and energy (E). (2) The canonical ensemble (NVT) in which N, V and temperature (T) are held constant. (3) The NVTp ensemble, an NVT with a pressure (P) bath giving an apparent constant pressure. The NVTp ensemble is an approximation to the Isothermal-Isobaric NPT ensemble.

K. DRP (Dynamic Reaction Path)

The dynamic reaction path[36] (DRP) method is a classical trajectory approach, which is based on quantum chemical energies that are computed "on-the-fly". There is no need for an *a priori* knowledge of the potential energy surface (PES). Energy is strictly conserved along the dynamic reaction path; therefore large step sizes are feasible. The initial dynamic reaction path method used the vibrational normal modes within GAMESS.

By selecting specific modes and adding kinetic energy (in units of quanta) to one or more modes one may direct the reaction to desired products. However, mode-mode mixing often occurs, allowing energy that is initially applied to one mode to leak into other modes. Thus, the amount of energy added should be in excess of the reaction/decomposition barrier in order to induce a reaction. The DRP method is a direct dynamics method that uses *ab initio* gradients/forces for each step which in turn are used to solve Newton's equations of motion and propagate the system

L. FMO (Fragment Molecular Orbital)

The fragment molecular orbital (FMO) method[37-40] was designed to reduce the scaling of fully QM methods by subdividing a system into fragments, while using a Coulomb

operator to treat the long-range interactions of the full system. The fragments in the FMO method are created electrostatically as shown in Figure 2. The FMO method assigns both electrons of the bonding pair to one fragment and none to the other. To avoid charged species, the deficient fragment has a proton reassigned to the electron rich species. This creates two neutral fragments indicated by the 1 and 5 in Figure 2.

Each individual fragment (monomer) is calculated in the electrostatic potential (ESP) of the entire system. These monomers are iterated to self-consistency within the ESP. Many body effects are accounted for by the *ab inito* calculations of two fragments (dimers; called the FMO2 method) and three fragments (trimers; called the FMO3 method). The FMO method evaluates the energy of a system by first calculating the initial electron density distribution of each monomer in the Coulomb bath of the system. Then, the Fock operators for each monomer are created to calculate the monomer energies. The ESP is converged by iterating the monomer energies to self-consistency. The dimer and trimer energies are not iterated to self-consistency. The FMO expression for the total energy is given by:

$$E = \sum_I^N E_I + \sum_{I>J}^N (E_{IJ} - E_I - E_J)$$

$$+ \sum_{I>J>K}^N \{(E_{IJK} - E_I - E_J - E_K) - (E_{IJ} - E_I - E_J) \tag{40}$$

$$- (E_{JK} - E_J - E_K) - (E_{KI} - E_K - E_I)\} + ...$$

where E_I, E_{IJ}, and E_{IJk} are monomer, dimer, and trimer energies, respectively. Terms are subtracted to avoid double counting.

In order to capture solvent effects the FMO method has been interfaced with the effective fragment potential (EFP)[38] method and the polarizable continuum model (PCM).[39]

The ability to use multiple wavefunction types for individual fragments is available through the multi-layer FMO (MFMO) implimentation.[40] The FMO method allows the monomers to be described by a variety of methods (i.e., RHF, MP2, DFT, CC, MCSCF).

M. SFM (Systematic Fragmentation Metho)

The systematic fragmentation method[41-43] (SFM) divides a molecular system into fragments while significantly decreasing the computational expense. It retains nearly the accuracy of a full *ab initio* calculation at the same level of theory. The total energy of the larger system can be obtained through addition and subtraction of the contributions from overlapping sub-systems or "groups".

Within the context of the SFM, a molecule can be thought of as a collection of functional groups. For example, ethanol contains three functional groups (CH_3, CH_2, and OH) in this framework, as shown in Figure 3. To divide the system into functional groups, single bonds must be broken. This process yields two fragments that are each assigned an electron from the bonding pair of the broken bond. A "cap" (hydrogen atom) is applied to the dangling bonds that are created by the fragmentation to avoid the creation of radical species. This hydrogen cap points in the direction of the broken bond at a chemically reasonable distance. By design, double or triple bonds are not broken, and the relevant atoms are kept as a part of one functional group. For instance, after the addition of the hydrogen caps, the functional groups from ethanal, CH_3CHO, would then be CH_4 and CH_2O.

As the separation between the breaks is increased, the accuracy increases, because the larger fragments give a more accurate description of the full system. In the limit, there is no fragmentation at all. The separation between broken bonds can be described as different

"levels" of the SFM, and are defined as follows:[42]

Consider the molecule *M*:

$$M = G_1G_2G_3G_4G_5G_6G_7G_8 \tag{41}$$

In the Level 1 SFM, two bonds separated by just one functional group are sequentially broken. The initial fragments would be described as follows:

$$M \approx G_1G_2 + G_2G_3G_4G_5G_6G_7G_8 - G_2 \tag{42}$$

The G_2 fragment is subtracted in order to preserve the number of atoms in Molecule, M, and to avoid double counting. This procedure is repeated on the $G_2G_3G_4G_5G_6G_7G_8$ fragment until no term is larger than two functional groups. The total energy of molecule M at fragmentation level 1 can then be decomposed into the following sum:

$$E_{\text{level 1}}^{\text{bonded}}(M) = E(G_1G_2) + E(G_2G_3) + E(G_3G_4) + E(G_4G_5) + E(G_5G_6) + E(G_6G_7) + E(G_7G_8) \\ - E(G_2) \quad - E(G_3) \quad - E(G_4) \quad - E(G_5) \quad - E(G_6) \quad - E(G_7) \tag{43}$$

In the level 2 SFM, bonds separated by two groups are broken. The total energy of molecule M at fragmentation level 2 is as follows:

$$E_{\text{level 2}}^{\text{bonded}}(M) = E(G_1G_2G_3) + E(G_2G_3G_4) + E(G_3G_4G_5) - E(G_2G_3) - E(G_3G_4). \tag{44}$$

In the level 3 SFM, bonds separated by three functional groups are sequentially broken, giving rise to the following energy expression:

$$E_{\text{level 3}}^{\text{bonded}}(M) = E(G_1G_2G_3G_4) + E(G_2G_3G_4G_5) + E(G_3G_4G_5G_6) + E(G_4G_5G_6G_7) + E(G_5G_6G_7G_8) \\ - E(G_2G_3G_4) \quad - E(G_3G_4G_5) \quad - E(G_4G_5G_6) \quad - E(G_5G_6G_7) \tag{45}$$

To describe a system analogously to the *ab inito* full molecular system non-bonded interactions among separated functional groups much be considered. Within the SFM framework a modified many body expansion is employed,[43] which assumes the bonded

interactions are much stronger than non-bonded interactions. Consider an example of a three-body non-bonded interaction between three functional groups G_1, G_2 and G_3. Any interaction is assumed to be negligible unless two of the groups are bonded to each other. Consider G_3 bonded directly to G_2 where the three-body interaction energy would be:

$$E_{nb}^{(1,2)}[G_1;G_2,G_3] = E(G_1G_2G_3) - E(G_1) - E(G_2G_3)$$
$$- E_{nb}^{(1,1)}[G_1;G_2] - E_{nb}^{(1,1)}[G_1;G_3] \quad , \tag{46}$$

The total SFM energy of a system is simply the addition of the bonded and non-bonded energies,

$$E_{SFM}^{total} = E^{bonded} + E^{non-bonded}. \tag{47}$$

The SFM is limited in the types of systems that may be considered. The SFM is unable to fragment conjugation in delocalized molecular systems, since breaks only occur between groups joined by single bonds. Secondly, SFM is unable to fragment six member rings using level 3 since the capping hydrogens would approach each other too closely and would therefore cause unphysical repulsive interactions. Therefore, a ring repair rule is imposed in which the ring itself is taken to be a functional group. Such ring repair rules extend to five and four membered rings for SFM levels 2 and 1 respectively. Previous work[41,42] has shown that level 3 with non-bonded interactions is needed to achieve high accuracy (1 kcal/mol) compared to fully *ab initio* calculations.

N. References

1. (a) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 361. (b) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 489. (c) Schrödinger, E. *Ann. Physik.* **1926**, *79*, 734. (d) Schrödinger, E. *Ann. Physik.* **1926**, *80*, 437. (e) Schrödinger, E. *Ann. Physik.* **1926**, *81*, 109.
2. Born, M.; Oppenheimer, R. *Ann. Physik.* **1927**, *84*, 457.

3. Hartree, D.R. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 426. (b) Hartree, D. R. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 111. (c) Hartree, D. R. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 89. (d) Fock, V. *Physik.* **1930**, *61*, 126.

4. (a) Hückel, E. *Z. Elektrochem.* **1937,** *43*, 752. (b) Hückel, E. *Z. Elektrochem.* **1937,** *43*, 827.

5. Roothan, C. C. *J. Rev. Mod. Phys.* **1960**, *32*, 179.

6. Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*, **1989**.

7. Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *B864*, 136.

8. Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Let.* **2003**, *91*, 146401.

9. Becke, A.D. *J. Chem. Phys.* **1993**, *98*, 5642. (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

10. Bartlett, R. J. *Ann. Rev. Phys. Chem.* **1981**, *32*, 359. (b) Binkley, J. S.; Pople, J. A. *Int. J. Quantum Chem.* **1975**, *9*, 229.

11. Møller, C.; Plesset, S. *Phys. Rev.* **1934**, *46*, 618.

12. Paldus, J. *NATO ASI Series, Series B: Physics* **1992,** *293*, 99. (b) Bartlett, R. J. *Adv. Ser. Phys. Chem.* **1995**, *2*, 1047. (c) Lee, T. J.; Scuseria, G. E. *Understanding Chemical Reactivity* **1995**, *13*, 47. (d) Piecuch, P.; Kucharski, S. A.; Kowalski, K.; Musial, M. *Com. Phys. Com.* **2002**, *149*, 71

13. Lennard-Jones, J. E. *Proc. R. Soc. London. A* **1924**, *106*, 463.

14. Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.

15. Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. J. Chem. Phys. **2006**, *124*, 104101.

16. Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J-P. *J. Chem. Theory. Comput.* **2007**, *3*, 1960.

17. (a)Xie, W.; Gao, J. *J. Chem. Theor. Comput.* **2007**, *3*, 1890. (b) Xie, W.; Pu, J.; MacKerell, A. D., Jr.; Gao, J. *J. Chem. Theor. Comput.* **2007**, *3*, 1878. (c) Xie, W.; Song, L.; Truhlar, D.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108.

18. Day, P.N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W.J.; Krauss, M; Grammer, D.; Basch H.; Cohen, D. J. *J. Chem Phys.* **1996**, *105*, 196

19. Adamovic, I.; Freitag, M.A.; Gordon, M. S. *J. Chem. Phys.* **2003**, *118*, 6275

20. Song, J.; Gorodon. M.S.; (unpublished)

21. Gordon, M.S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. *J. Phys. Chem. A* **2001**, *105*, 293.

22. (a) Stone, A. J. The Theory of Intermolecular Forces, **1996.** (b) Stone, A. J. Chem. Phys. Lett. **1981**, 83, 233

23. Slipchenko, L.; Gordon, M.S. *J. Comp. Chem.* **2006**, *28*, 276.(b) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. *J. Chem. Phys.* **2000**, *112*, 7300

24. Li, H.; Gordon, M.S., *Theor. Chem. Accts.* **2006**, *115*, 385. (b) Li, H.; Gordon, M.S.; Jensen, J.H. *J. Chem. Phys.* **2006**, *124*, 214107.

25. Tomasi, J.; Cammi, R.; Mennucci, B.; Cappelli, C.; Corni, S. *Phys. Chem. Chem.Phys.* **2002**, *4*, 5697. (b) Luque, F. J.; Curutchet, C.; Munoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpi, J. L.; Orozco, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3827. (c) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, 99, 2161-2200. (d) Truong, T. N. *Intern. Rev. Phys. Chem.* **1998**, *17*, 525.

26. Miertus, S.; Scrocco, E.; Tomasi, J. Chem.Phys. **1981**, 55, 117-129. (b) Cances, E.; Mennucci, B.; Tomasi, J. *J. of Chem. Phys*. **1997**, *107*, 3032.
27. Li, H.; Pomelli, C.; Jensen, J. *Theo. Chem. Acta*. **2003**, *109*, 71.
28. Klamt, A.; Schuurmann, G. *J. Chem. Soc.-Perkin Trans*. **1993**, *2*, 799. (b) Curutchet, C.; Orozco, M.; Luque, F. J. *J. Comp. Chem*. **2001**, *22*, 1180. (c) Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Let*. **1998**, *288*, 293. (d) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys*. **1997**, *107*, 3210.
29. Stefanovich, E. V.; Truong, T. N. *Chem. Phys. Let*. **1995**, *244*, 65. (b) Pascualahuir, J. L.; Silla, E.; Tunon, I. *J. Comp. Chem*. **1994**, *15*, 1127. (c) Silla, E.; Tunon, I.; Pascualahuir, J. L. *J. Comp. Chem.* **1991**, *12*, 1077. (d) Silla, E.; Villar, F.; Nilsson, O.; Pascualahuir, J. L.; Tapia, O. *J. Mol. Grap*. **1990**, *8*, 168-172. (e) Pascualahuir, J. L.; Silla, E. *J. Comp. Chem*. **1990**, *11*, 1047.
30. Metropolis, N.; Rosenbluth, A.; Teller, A. *J. Chem. Phys*. **1953**, *21*, 1087.
31. Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys*. **2000**, *112*, 2063.
32. Hockney, R. W. *Meth. Comp. Phys*. **1970**, *9*, 136.
33. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys*. **1982**, *76*, 637.
34. (a) Verlet, L. *Phys. Rev*. **1967**, *159*, 98. (b) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*. Oxford University Press, USA, **1989**.
35. Dahlquist, G.; Björk, A. *Numerical Methods*. Prentice Hall. Englewood Clifs NJ. [3.2.1] **1974**.
36. Taketsugu, T.; Gordon, M. S. *J. Chem. Phys*. **1995**, *103*, 10042.
37. (a) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett*. **1999**, *312*, 319. (b) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701. (c) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett*. **2000**, *318*, 614. (d) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett*. **2001**, *336*, 163. (e) Inadomi, Y.; Nakano, T.; Kitaura, K.; Nagashima, U. *Chem. Phys. Lett*. **2002**, *364*, 139. (f) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett*. **2002**, *351*, 475.
38. Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.
39. Mochizuki, Y.; Tanaka, K.; Yamashita, K.; Ishikawa, T.; Nakano, T.; Amari, S.; Segawa, K.; Murase, T.; Tokiwa, H.; Sakurai, M. *Theor. Chem. Acc*. **2007**, *117*, 541.
40. Goto, H.; Obata, S.; Kamakura, T.; Nakayama, N.; Sato, M.; Nakajima, Y.; Nagashima, U.; Watanabe, T.; Inadomi, Y.; Ito, M.; Nishikawa, T.; Nakano, T.; Nilsson, L.; Tanaka, S.; Fukuzawa, K.; Inagaki, Y.; Hamada, M.; Chuman, H. *In Modern Methods for Theoretical Physical Chemistry and Biopolymers*; Starikov, E. B., Lewis, J. P., Tanaka, S., Eds.; Elsevier: Amsterdam, **2006**; pp 227-248.
41. Deev, V.; Collins, M.A. *J. Chem. Phys*. **2005**, *122*, 154102. b) Collins, M.A. *J. Chem. Phys*. **2007**, *127*, 024104.
42. Collins, M A; Deev V. A. *J. Chem. Phys*. **2006**, *125*, 104104.
43. Netzloff, HM; Collins, M.A. *J. Chem. Phys*. **2007**, *127*, 134113.

Figure 1. This pictorial representation[53c] shows the original Verlet method[53d] in a), the leapfrog method in b), and the velocity Verlet method in c). The position(r), velocity(v) and acceleration(a) are listed along the left hand side, while the time(t) and time steps ($\pm\delta t$) are on the top of each step(9x9 box) of the algorithm. The offset boxes are at a $1/2\delta t$ time step. The red boxes represent stored variables and the arrows illustrate the use of stored data at their origin to solve for the variable at their termination. This is done via Eqs. (37)-(39) for the respective method, and $\ddot{r}(t) = a(t)$.
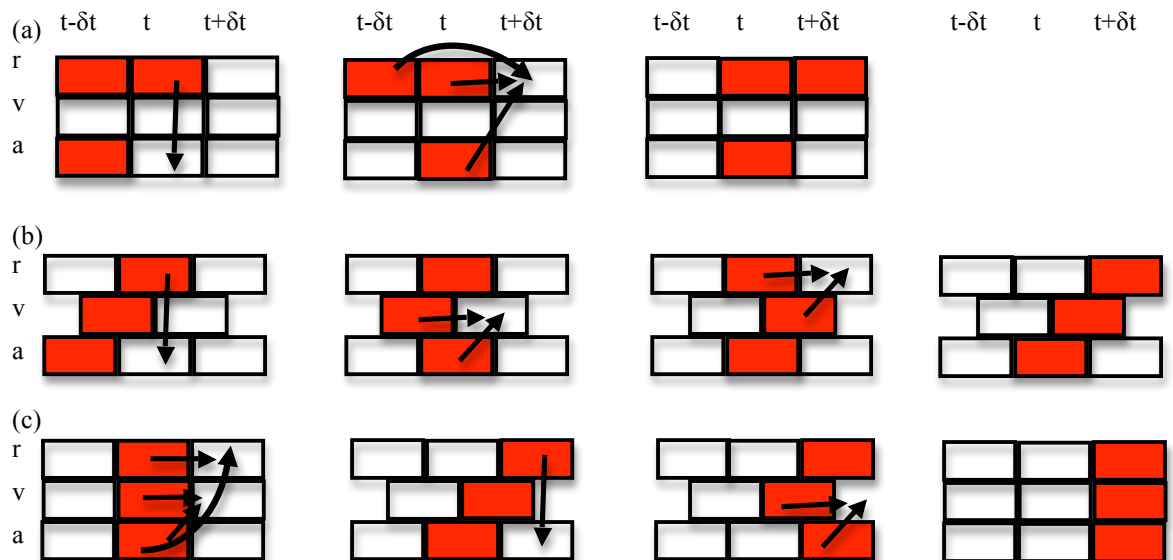
Figure 2. Electrostatic (heterolytic) fractionation of a bond in the FMO method.
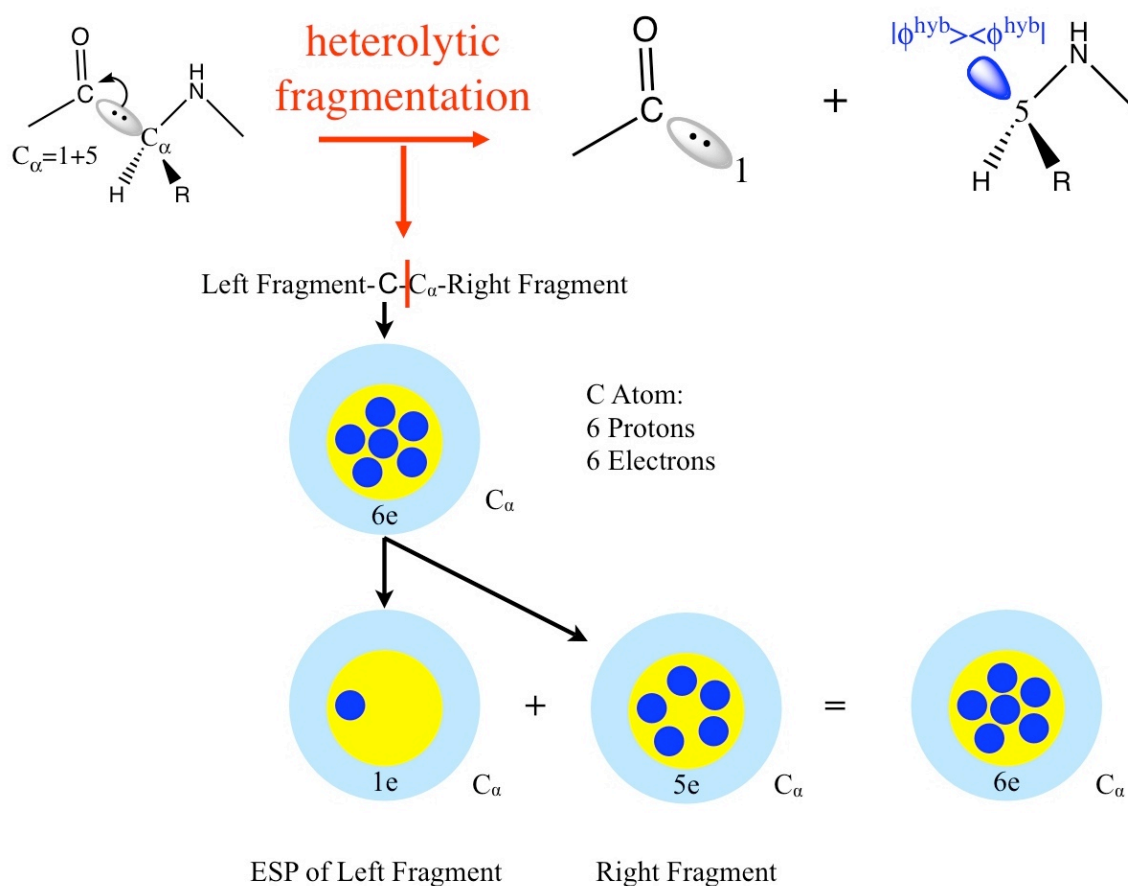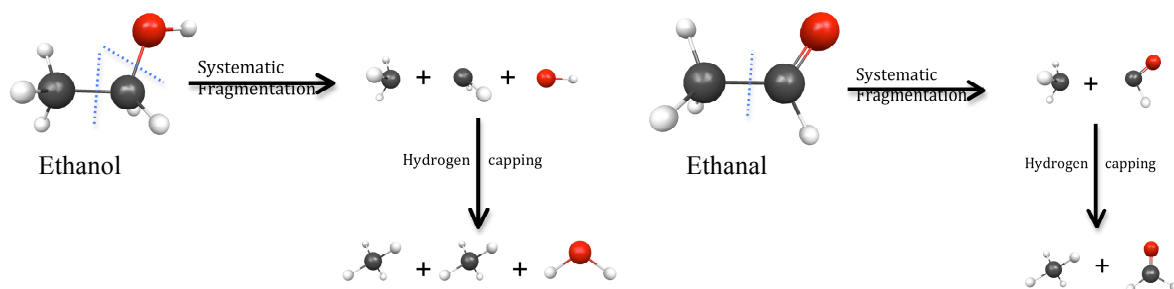
Figure 3. Pictorial examples of level one fragmentation for ethanol and ethanal. The first step breaks bonds creating functional groups, with hydrogen caps.

ALANINE: THEN THERE WAS WATER

Jonathan M. Mullin and Mark S. Gordon*

Department of Chemistry, Iowa State University, Ames, Iowa 50011

*mark@si.msg.chem.iastate.edu

Abstract.

An *ab initio* study of the addition of successive water molecules to the amino acid L-alanine in both the nonionized (N) and zwitterionic (Z) forms are presented. The main focus is the number of waters needed to stabilize the Z form, and how the solvent affects conformational preference. The solvent is modeled by *ab initio* electronic structure theory, the EFP (effective fragment potential) model, and the isotropic dielectric PCM (polarizable continuum method) bulk solvation techniques. The EFP discrete solvation model is used with a Monte Carlo algorithm to sample the configuration space to find the global minimum. Bridging structures are predicted to be the lowest energy Z minima after 3-5 discrete waters are included in the calculations, depending on the level of theory. Second-order perturbation theory and PCM stabilize the Z structures by ~3-6 kcal/mol and 7 kcal/mol, respectively, relative to the N global minimum through the addition of up to 8 waters. Subsequently, the contributions of each are ~1 kcal/mol relative to the N global minimum. The presence of 32 waters appears to be close to converging the N-Z enthalpy difference, ΔH_{N-Z}.

Introduction

Alanine is one of the 20 amino acids that are the naturally occurring, basic building blocks of proteins. A basic understanding of phenomena associated with alanine and other amino acids is essential to understanding the complexity of protein folding and protein interactions and reactivity. Modeling of alanine in solution is useful for exploring the flexibility of small amino acids, so both glycine and alanine have been extensively studied, both experimentally and theoretically.[1-44] Amino acids in the gas phase are predominantly in the nonionized (N) form; however, when placed in aqueous solution the zwitterionic (Z) form is generally more stable.[1-4,17,39,43]

The study of amino acid solvation can employ various methods, both discrete (explicit) and continuum (implicit). Previous studies that used continuum models treated the solvent as a polarizable medium. The most commonly employed continuum methods are based on the polarizable continuum method (PCM),[46-49] self consistent reaction field (SCRF) or Onsager cavity model,[50] solvation model 8 (SM8),[51] COSMO,[52] GCOSMO,[53] and solvation with volume polarization (SVP).[54] Advantages of the continuum approach are computational efficiency, simplicity of the calculations, and prediction of bulk properties with reasonable accuracy. However, these methods are sensitive to the parameterization of the size and shape of the cavity that surrounds the solute. Further, continuum models do not always describe important solute-solvent electronic effects well. The continuum models cannot describe systems that form hydrogen bonds between solute and solvent accurately, since they do not include explicit solvent-solute interactions.

The discrete methods account for intermolecular solute-solvent interactions. Common implementations include *ab initio* quantum mechanics (QM), which rapidly becomes too computationally demanding with increasing numbers of solvent molecules,

and hybrid quantum mechanics/molecular mechanics (QM/MM) methods. QM/MM methods reduce the computational demands by orders of magnitude.

One complication that arises when using discrete solvents is the rapidly increasing number of geometric degrees of freedom and the even more rapidly increasing number of configurations that may be important in configuration space. A molecular dynamics (MD) or Monte Carlo (MC) method must be used to adequately sample and accurately capture the full configuration space.

A promising QM/MM method is the effective fragment potential (EFP) approach.[55-60] The original, Hartree-Fock based, EFP1 method was designed specifically for water and is represented by a set of one-electron potentials that are added to the *ab initio* electronic Hamiltonian. The EFP1 method contains three energy terms: Coulomb, induction/polarization and a remainder term that includes exchange repulsion and charge transfer for both solvent - solvent and solute - solvent interactions.

The Coulomb portion of the electrostatic interaction is obtained using Stone's distributed multipolar analysis,[60b] truncated at the octopole term. Atom centers and the bond midpoints are used as expansion points.

Induction (polarization) is the interaction of an induced dipole on one fragment with the permanent dipole on another fragment, expressed in terms of the dipole polarizability. The efficacy of truncating the polarizability expansion at the first (dipole) term is due to the EFP treatment of this term in a distributed manner: The molecular polarizability is expressed as a tensor sum of localized molecular orbital (LMO) polarizabilities. Therefore, the number of bonds and lone pairs in the molecule gives the number of polarizability points. Iterating the dipole-induced dipole interaction to self-consistency captures

many body effects.

Charge penetration (damping of the Coulomb term) is included in order to account for short-range quantum effects that are not accounted for by the classical multipolar expansion. The first two EFP terms are determined based on QM calculations on the water monomer. The third term is fit to a quantum mechanical water dimer potential. Three EFP1 methods have been derived, based on Hartree-Fock (EFP1/HF), density functional theory (EFP1/DFT),[61] which includes short range correlation effects, and second order Møller-Plesset perturbation theory (EFP1/MP2),[62] which includes dispersion effects (a fourth term) and a second order correction to the Coulomb and polarization terms. EFP1/MP2 is currently only available for solvent-solvent, not solvent-solute, interactions. A general effective fragment potential model (EFP2) that contains no empirically fitted parameters has also been developed.[55-57] An EFP2 can therefore be generated for any molecular species. In addition to the Coulomb plus damping and polarization/induction terms described above for EFP1, the EFP2 model[63] includes interaction terms that describe exchange repulsion, dispersion,[64] and charge transfer,[65] each of which have been derived from first principles. The EFP2 method has been described in detail in a recent review.[63] Both EFP1 and EFP2 have been used in the present study.

Some three-layer approaches have been developed that combine the discrete and continuum approaches.[21] These methods may reduce the number of explicit solvent molecules required (and thus the computational cost) and may facilitate an accurate description of long-range interactions.

Interactions between amino acids and water molecules are of considerable interest, particularly with regard to structures and vibrational spectra. Depending on the pH,

amino acids may be N, Z, anionic, or cationic. The intermolecular hydrogen bonding interaction that occurs in water leads to a considerably larger stability of the Z form compared to the N form. Consequently, the experimental free energy and enthalpy for the process $Z_{(aq)} \rightarrow N_{(aq)}$ are 7.3 and 10.3 kcal/mol, respectively, for glycine.[66] The corresponding experimental values for alanine do not appear to be available; however, they are likely to be similar to the glycine values.

Most previous studies of the Z/N equilibrium in amino acids have focused on glycine,[1-27] the simplest amino acid. The present study addresses similar questions, with regard to relative energies and structures, for alanine. Alanine (Figures 1-3) is the simplest amino acid to exhibit chirality because of its substituent methyl group. The crystal structure of alanine exhibits the zwitterion.[67,68] This structure is often used as a basis for the aqueous solvated geometry, as it is generally accepted that for small aqueous proteins, crystallization gives rise to just a small perturbation of the solution structure.[69-71] This stems mainly from the success of subsequent experimentation based on data from crystal structures. It should be noted that this observation implies that the Z structure is most likely dominant in aqueous solution, but not necessarily the same Z conformation that is observed in crystalline form.

Previous glycine studies

The earliest computational studies of amino acids focused on the relative energies of the N and Z structures of glycine in the gas phase[1,2,3] and in solution.[4] There are three internal rotational degrees of freedom for N glycine: rotation about the C-C bond, rotation of the hydroxyl group about the C-O bond and rotation of the amino group about the C-N bond. Initial studies focused on C_s and C_1 stationary points on the conformational

potential energy surface of gas phase N glycine.[5,6] If one uses a basis set that includes polarization functions on all atoms, the Z form is not a minimum on the glycine potential energy surface.[7] The lowest energy N species has a structure that is well positioned to accommodate an intermolecular proton transfer.[9,13] Consequently, in the gas phase the Z undergoes proton transfer to yield the N structure.

In order to establish the Z as a local minimum on the potential energy surface, electrostatic stabilization in the form of aqueous solvation must be included in the computation. Continuum methods combined with HF, DFT, and MP2 have predicted that the electrostatic solute-solvent interactions stabilize the Z form, so that the zwitterion does not undergo spontaneous hydrogen transfer.[8-15] Discrete solvent studies of glycine have primarily focused on a small number of water molecules. The recent paper by Aikens and Gordon presents an extensive review of these calculations.[17] A single water molecule will cause the Z isomer of glycine to become a local minimum at the RHF level of theory; however, when correlation is introduced, for example with MP2, this local minimum disappears. It appears to be necessary to have two explicit waters present in order to maintain a local Z minimum at the MP2 level of theory.[18] A continuation of the discrete method to a full solvent shell greatly increases the number of degrees of freedom in coordinate space. Thus, the number of local minima would increase to the point at which an automated sampling technique, such as Monte Carlo (MC) or molecular dynamics (MD) simulations, must be employed to efficiently sample this space. Alternatively, a three-layered approach has been used to reduce the number of discrete waters needed by including a continuum that surrounds the entire atomistic system. These methods may reduce the computational effort. However the long range interactions may not achieve the

desired accuracy.[16 I,14,16H,7,16G,15,19-24]

Supermolecular complexes have also been studied using DFT/B3LYP[21] and the composite DFT/B3LYP + Onsager[36] model, but these authors do not discuss configurational sampling. In a separate effort, MC simulations with a molecular mechanics force field were employed to sample the configuration space, with the structures subsequently optimized with semi-empirical QM methods.[19] This study found that the Z form is more stable than the N form when 7 and 15 waters are present in a self-consistent reaction field. Bandyopadhyay et al. considered the glycine$(H_2O)_8$ complex using three-layer models that combined the effective fragment potential with either the Onsager or PCM models.[16,21] In the first study, eight water molecules in the solvent shell were selected from a previous molecular dynamics calculation, and this structure was optimized.[21] In the second study, Monte Carlo simulations with local minimizations were used in order to find a low energy configuration.[16] Cui used a similar approach in order to examine the affects of TIP3P waters.[20] Using the three-layer approach, the Z form is predicted to be more stable than the N species when a reliable continuum method[21,20,16] is used. Campo et al, studied the effects of increasing glycine concentration, and separately the inclusion of Na^+ and Cl^- on the water orientation via MD.[27]

Previous Alanine Studies

Previous studies of alanine have examined the gas phase structures and found the energy difference to favor the N isomer at several levels of theory: AM1,[19,28] RHF,[29] DFT/B3LYP,[29,30,28] DFT/PWP,[29] MP2,[29] and coupled cluster with single and double excitations (CCSD).[31] The Z form was found to be a local minimum with B3LYP and MP2. Continuum studies using PCM and DFT/B3LYP predict the Z isomer to be lower in en-

ergy than the N species.[28] MC simulations and generalized Born (GB) studies of the Z conformation with 212 explicit waters predict a nearly planar NCCOO moiety,[32] for which the barrier to rotation of the COO is 5.9 kcal/mol and the NH_3^+ rotation barrier is less than 1 kcal/mol. The experimental observations show free rotation NH_3^+ of glycine in spectroscopy of water-amino acid microjets.[33]

Discrete solvation with a small number of waters has also been explored. The N (H_2O) complex has been explored with DFT/B3LYP, using the 6-311++G(d,p) basis set.[34] Kwon et al. found large barriers to proton transfer from N to Z, but a 0.85 kcal/mol barrier with MP2 for the reverse process when two waters are present. The transition state for this process lies near the Z isomer on the potential energy surface.[35] Ahn et al. explored the one and two water complexes and found a concerted double proton transfer for the two water case with B3LYP and MP2.[36] Park et al. determined a 5.89 kcal/mol barrier for a concerted triple proton transfer mechanism using MP2/6-31++G(d,p) with three water molecules present to connect the local minima of the Z and N isomers.[37] Xu et al. showed that alanine in the presence of 7-9 waters has high electron binding energy spectral trails, which implies that the Z species has been formed by the addition of 7 waters.[38] Chuchev et al. used DFT/B3LYP and MD to systematically study the preference for Z versus N for 1-10 waters, and proposed that the conversion to the Z form occurs in the range of 6-8 waters.[39]

Ellzy computed structures of alanine with four waters and compared them to vibrational circular dichroism data.[40] Ellzy's lowest energy structure is not the global minimum found by Chuchev. However, agreement with vibrational circular dichroism does imply that there is a population of an approximately energetically degenerate state

near the minimum. This suggests that Boltzmann averaging may be useful in interpreting computational data.[40]

Use of continuum models in a three-layer approach employing the Onsager reaction field and DFT/B3LYP waters[41,42] concluded that 9 waters was not a complete solvent shell. Explicit hydration around the methyl group was not observed when the continuum was employed, and the explicit water model did not reproduce the experimental vibrational absorption spectrum. A large number of discrete TIP3P waters has also been used with MD to examine alanine. This study concluded that a large number of water molecules is needed to make solvation of the methyl group energetically favorable.[43] These simulations contained 300 rigid water molecules and exhibited a smaller probability of water density near the methyl group. Dixit et al. examined the free energies of hydration of amino acids. They predicted that the pK_a of glycine and alanine are similar to each other, based on a simple electrostatic model used to calculate the pK_a shifts. This suggest that the ΔH_{vap} of alanine may be similar to that of glycine ($\Delta H=10.2$ kcal/mol).[44]

The present work examines the crossover from the N to the Z form of alanine as the number of water molecules is increased. MC with simulated annealing was employed to find local minima and to attempt to determine the global minimum. Quantitatively, it is of interest to determine how many water molecules are necessary to converge the Z-N alanine energy difference in aqueous solution. Both discrete and continuum solvent models have been used to elucidate the effectiveness of both approaches.

Computational methods

The two major forms of alanine explored here are the N (Figures 1 and 2) and Z (Figure 3) isomers. Starting structures for alanine were obtained from the crystal struc-

ture,[67,68] with additional structures exploring torsions about the COO^-, $COOH$ and NH_3^+, NH_2 (Z,N) units, as well as the chirality about the alpha carbon. Other starting N structures allowed for rotation of the OH bond away from or toward the NH_2 or CH_3 group, depending on the torsion of the COOH group. The initial gas phase results for both chiral forms of N alanine (L, D) were obtained with restricted Hartree-Fock[72] (RHF) and Møller-Plesset second order perturbation theory[73] (MP2) using the 6-31++G(d,p) basis set.[74-76] In order for the Z structures to be stable with no discrete water molecules present, a continuum solvent model, such as PCM, must be used. The PCM uses van der Wall radii for cavity generation. The values used in these calculations may be found in the GAMESS manual. Therefore, the Z structures were found using RHF in PCM (RHF+PCM), employing the 6-31++G(d,p) basis set. The N structures were also re-optimized using RHF+PCM/6-31++G(d,p). A MP2+PCM single point was run at the RHF+PCM structure for both the N and Z forms.

Once the gas phase (plus PCM for Z) structures were determined, a discrete approach was employed to examine the effect of systematically adding waters. First, alanine(H$_2$O)$_n$, n = 1-3, clusters were optimized at the RHF and MP2 levels of theory with the 6-31++G(d,p) basis set. The 1-3 waters were added manually in chemically sensible orientations. Each water molecule was modeled by both *ab initio* QM and EFP1/HF when using HF for alanine, and by EFP1/DFT water when using MP2 to describe alanine. Optimizations with both water and alanine treated as MP2 allowed the testing of possible hydrogen transfer.

For larger numbers of waters, manually choosing their placement is not efficient or effective, so the Monte Carlo[77,78] method with simulated annealing (SA) was used to

sample configuration space. The MC simulations were carried out using EFP2 for both the solute and the solvent. The molecular structures used to create the EFP2s were taken from RHF+PCM calculations. All eight forms of L-alanine were explored along with ALA 1D and ALA 2D. For the Z form, the L enantiomer was considered, as well as three Z rotamers (45˚ and 90˚ rotation of the COO⁻ group about the C-C bond, 60˚ rotation of NH_3^+ about the N-C bond). The MC/SA method with local minimization was used to sample the configuration space. For each global minimum found, the number of structures sampled was on the order of 400,000 – 1,000,000. A local minimization was performed every 10 or 100 steps, and the number of steps taken for each temperature was varied (e.g., 100, 500, 1000, 10,000). The number of fragments moved per step was also varied in the calculations. With six or more discrete solvent molecules present, one to five fragments were moved during each successive step. The starting temperature for the simulated annealing runs varied from 300 to 20,000 K and the final temperature was kept at 200 K. The MC/SA code was also used to verify the minima found for one to three waters. Both approaches found the same structures (except when MP2 predicted proton transfer) through three solvent molecules, at which point the MC technique was deemed effective and used exclusively.

The twenty lowest unique minima were selected for further investigation for each rotamer of alanine explored with n waters. Optimizations and Hessians (matrices of energy second derivatives) were calculated at the EFP2 level of theory. Further, these structures were evaluated by MP2 optimizations and Hessians to assess the accuracy of EFP2. Single point calculations were performed using QM alanine with water represented by (EFP1/HF) or (EFP1/DFT). EFP1 employs a set of potentials that are added to the one-

electron part of the QM electronic Hamiltonian.

The notation employed to indicate the type of methods used for the single point energies is alanine(water)//geometry. For example, HF(EFP1/HF)//EFP2 represents a single point energy calculation with the alanine treated as RHF and the water as EFP1/HF calculated using an EFP2 geometry. Single point energies were calculated using HF(EFP1/HF)//EFP2, DFT(EFP1/DFT)//EFP2 using the B3LYP functional,[79] MP2(EFP1/DFT)//EFP2, and MP2(MP2)//EFP2. Optimizations starting at the EFP2 geometries used MP2(EFP1/DFT), where alanine is MP2 and the water is (EFP1/DFT). Another set of single points were run using MP2(MP2)//MP2(EFP1/DFT), where the both the alanine and the water were treated as MP2. Single points including C-PCM+MP2 were run at the EFP2 and MP2(EFP1/DFT) optimized geometries. All calculations were performed with the electronic structure code GAMESS,[80] which is freely available from Iowa State University at http://www.msg.chem.iastate.edu/. Structures were visualized with MacMolPlot,[81] a graphical interface to GAMESS that is available at the same web site.

Results

Alanine

The sixteen N alanine conformations found are shown in Figure 1(L chirality) and Figure 2 (D chirality). All L structures along with ALA 1 D and ALA 2 D, (see Figures 1, 2) were used in the MC simulations. The EFP2 Z form and the conformational rotamers (45˚, 90˚ rotation of the COO$^-$ group about the C-C bond, 60˚ rotation of NH$_3^+$ group about the N-C bond) used in the MC/SA simulations are in Figure 3.

L-Alanine is the biologically active form of the molecule, so it is the primary fo-

cus of the present study. No significant difference in energy due to chirality was found, as can be seen in Table 1. The origin of the HF and MP2 relative energy differences for the ALA8 L and ALA8 D isomers (Table 1) stems from a small difference in the methyl orientation relative to the N-C-C-O plane. While most of the data reported relates to L-alanine the N species of D-alanine with 3-8 waters was also studied. Table 2 shows that with PCM+MP2 the L vs. D energy differences are within ~1 kcal/mol. These differences may be the result of incomplete sampling of the D-alanine configuration space, because only two D-alanine conformers were used, and the number of energy evaluations was not as large.

Continuum Calculations

The HF+PCM continuum method predicts that the Z form is lower in energy than the N species; however, this difference is only 0.8 kcal/mol. While the HF energy difference is qualitatively correct, the importance of including electron correlation has been demonstrated in previous studies.[5,9] A MP2+PCM single point calculation predicts the Z form to be 1.8 kcal/mol lower in energy. Thus, MP2 provides an additional 1 kcal/mol stabilization of the Z form relative to the N form. MP2 may alter the energy gradient, and consequently, the geometry, as well. Furthermore, the strengths of the hydrogen bonds formed between the solvent molecules and the N and Z isomers differ, and the continuum solvent may not fully capture this differential effect. A combined discrete+continuum approach might result in a faster convergence of the Z-N energy difference.

In the following subsections the impact of an increasing number of waters on the N-Z equilibrium and the detailed structures is examined. These structures are presented in

Figures 4-12. The ordering of the structures in these figures is determined by the relative energies of the MP2+PCM calculations, since MP2+PCM is considered to be the highest level of theory in this work.

Alanine(H₂O)₁ ₋ ₂ Figures 4, 5

Water may interact with a given alanine conformer at the carboxyl group, the amine group, or the saturated carbon. With one and two waters, in the majority of the low energy N (N1-1, N2-1) structures, the water molecules interact predominately with the carboxyl group which has been seen previously.[82] The lowest energy Z form, Z1-1, has a water bridge between the COO^- and NH_3^+ groups, even in the one water case, as shown in Figure 4.

At the HF level of theory, for both one and two water molecules, the Z form corresponds to a local minimum on the alanine-$(H_2O)_n$ potential energy surface; however, when electron correlation is added via MP2, the Z structures collapse via a proton transfer from the NH_3^+ to the COO^- group to form the N structure. MP2 optimizations of the N form do not significantly change the ordering of the sampled structures. There is some relaxation of the alanine structure from a frozen EFP2 geometry when the geometry is allowed to optimize with MP2(EFP1/DFT). However, on average, this does not create a significant change in the energy relative to the global minimum (~0.3 kcal/mol for the N conformer). N1-1 (Figure 4) is predicted to be the global minimum by 2.1 kcal/mol with one water and N2-1 (Figure 5) by 1.8 kcal/mol for two waters.

MP2 stabilizes the Z form relative to the N form, and MP2(EFP1/DFT) re-optimization further stabilizes the Z form. However, with 1-2 waters the Z form is predicted to be 8-10 kcal/mol higher in energy unless a continuum is also included. Note that

for both one and two waters, the Z form would collapse to the N form if the entire structure were fully optimized using MP2 without a continuum.

Alanine(H_2O)_3 Figure 6

For the N form, water binds in three distinct ways: separated, chain and carboxyl cluster. The separated arrangement in N3-1 has two waters bound to the carboxyl group that form a cyclic structure, while the third water binds to the amine group and does not form hydrogen bonds to the other two waters. The arrangement in N3-2 is a chain of hydrogen bonded waters connecting the carboxyl and amine groups. The carboxyl cluster arrangement is illustrated by N3-3 and N3-4. These carboxyl clusters, which resemble small water clusters, are energetically favorable except when including a continuum solvent.

At the MP2 level of theory (for both alanine and water), the lowest energy Z isomer is now stable and does not undergo spontaneous proton transfer. Since the Z isomer is a local minimum on the potential energy surface when three waters are present, the geometry obtained with the EFP2 model potential (which cannot treat proton transfers) should be qualitatively correct. In the lowest energy Z form, Z3-1, two of the waters interact directly with the carboxyl group, while the third water hydrogen bonds to the first two waters and the amino group. This third water is partially obscured by an oxygen in the lower right corner of Z3-1. The Z3-1 isomer is the global minimum only when both PCM and MP2 are included in the calculation. The higher energy Z species Z3-2 is lower in energy until both PCM and MP2 are included.

Alanine(H_2O)_4 Figure 7

For four or more waters, the global minimum structure is determined exclusively

using MC simulations, because of the large number of configurations that are required to properly sample the configuration space. For four waters, all methods that do not include PCM predict the lowest energy structure to be the neutral species N4-4. This isomer resembles the water hexamer "prism" structure, in which the two oxygen atoms in the COOH group act as two of the six centers. When PCM and MP2 are both included in the calculation, N4-1, similar in structure to the water hexamer "bag" structure, is found to be the lowest energy N structure, and the global minimum is now Z4-1. At the RHF/6-31++G(d,p) level of theory, Z4-1 is predicted to be 13.5 kcal/mol higher in energy than N4-4. The inclusion of dynamic electron correlation via MP2 at the same EFP2 geometry changes the energy difference to 8.7 kcal/mol. PCM+MP2//MP2(H2ODFT) single point energy calculations predict that Z4-1 is 1.8 kcal/mol lower in energy than N4-1 and 10.5 kcal/mol lower in energy than N4-4.

Alanine(H$_2$O)$_5$ Figure 8

In the majority of the N isomers with five water molecules present, one water molecule interacts with N alanine near the amine, while most of the waters interact via carboxyl clustering. N5-3 is an example of a structure in which the water interacts exclusively with the carboxyl group. As the level of theory is improved, there are noticeable changes in the energy ordering of the lowest energy isomers. This is especially true when both PCM and MP2 are included in the calculation with the MP2(EFP1/DFT) geometry. At this highest level of theory, Z5-1 is the global minimum, and Z5-2 (3.9 kcal/mol higher in energy than Z5-1) is also lower in energy than any N isomer. In these two Z isomers, the water molecules interact strongly with both alanine charge centers, NH_3^+ and COO^-. This leads to the formation of water bridges that connect the two charge centers,

as noticed previously for glycine.[17] The lowest energy neutral species, N5-1, has a chain of waters that connects the amino group with the carboxyl group, whereas the five waters cluster around the carboxyl group in N5-4. For five waters, all methods that include only discrete solvent molecules (no PCM) still predict the N form to be the global minimum.

Alanine(H₂O)₆ Figure 9

The pattern that emerged for four and five water molecules continues for six waters. In calculations that do not include PCM, the lowest energy species is still a N isomer, although the energy difference between the lowest energy N and Z isomers decreases as the number of waters increases. For example, at the MP2 (EFP1/DFT) - 31++G(d,p) level of theory, N6-4 is only 0.9 kcal/mol lower in energy than Z6-2. As noted above, when both MP2 and PCM are included in the calculation, there is a dramatic re-ordering of relative energies, with the global minimum now being Z6-1, followed in energy by Z6-2 < N6-1 < N6-2. The two lowest energy Z isomers, Z6-1 and Z6-2, exhibit bridging chains that connect the two charge centers. The addition of PCM greatly stabilizes Z6-1 relative to Z6-2, while MP2 has only a small effect.

Alanine(H₂O)₇ Figure 10

Some dramatic changes emerge for n = 7. At this point, all levels of theory except HF predict that the lowest energy structure is a zwitterion, even without the continuum present. PCM does stabilize Z7-1 relative to Z7-2 by about 5 kcal/mol, but Z7-1 is still the global minimum before the addition of PCM to the calculation. Interestingly, once electron correlation has been incorporated via MP2, the subsequent addition of PCM lowers the energy of the four N species shown in Figure 10 relative to Z7-1. This global minimum exhibits a chain of water molecules that connects the COO⁻ group with the

NH_3^+ group. In the lowest energy N isomer, N7-1, the seven water molecules are mostly clustered around the carboxyl group. The higher energy (by 4.4 kcal/mol) N7-3 species exhibits a greater propensity for waters to move near the amino group.

Alanine(H₂O)₈ Figure 11

The two lowest energy eight-water zwitterions are Z8-1 and Z8-2, with the former being lower in energy than the latter by 2.4 kcal/mol. These are the two lowest energy structures of the six (including four N isomers) shown in Figure 11. In the four N structures the carboxyl clustering motif dominates with some small water density near the amino group.

There is a trend thus far for the N structures to resemble small water clusters in which the carboxyl group tends to play the role of two waters in an extended water cluster. The lowest energy N minima consequently appear to be very similar to the global minima of small water clusters. This tendency is more apparent when the PCM contribution is omitted, in which case the preferred structures have the explicit water molecules saturating the hydrogen bonding amine and carboxyl groups. For 5-8 water molecules, the waters tend to form clusters among themselves, most likely due to stronger water-water hydrogen bonds. In the Z isomers, the presence of local charges strengthens the water-alanine interactions, and the waters form bridges or chains that connect the positive and negative charge centers. Up to eight water molecules, the waters have largely avoided the methyl group.

In the following paragraphs, the numbers of water molecules included in the calculation are increased in units of eight waters, in an attempt to approach "complete" solvation of alanine.

Alanine(H₂O)₁₆ Figure 12

When 16 waters are present, Z16-1 is predicted to be the global minimum at all levels of theory, and all four N isomers shown in Figure 12 are higher in energy than the two lowest energy Z isomers. In the lowest energy Z structures, the water molecules are well distributed along the $COO^-NH_3^+$ corridor, but they mostly avoid the hydrocarbon moiety. In the N structures, carboxyl clustering again dominates in all structures shown in Figure 12, suggesting that, as noted above, the water-water hydrogen bonding interaction is dominant. The Z form clearly causes more reorganization of the water than the N form when compared to pure water clusters. For the sixteen water case, it appears that the MP2(EFP1/DFT) optimization has a large effect, as seen in the relative energies of Z16-1 and N16-1 – 3.8 kcal/mol at EFP2 geometries using MP2(EFP1/DFT) single point energies. When the geometries are re-optimized with MP2(EFP1/DFT)/6-31++G(d,p), this relative energy increases to 15.6 kcal/mol. Single point calculations at the MP2(EFP1/DFT)/ 6-31++G(d,p) geometry, including PCM, lowers the energy of N relative to Z to 5.2 kcal/mol. These large relative energy changes occur for all of the structures shown in Figure 12. However, the two Z isomers are the lowest-energy 16-water species and N16-1 is the lowest energy N isomer at all levels of theory.

Alanine(H₂O)₂₄ Figure 13

For most of the 24-water Z isomers, the water molecules form a concave shape in which the alanine is embedded, although not fully solvated. The N isomers have the appearance of a water droplet, in which the alanine is integrated into the surface of the cluster, with both the amine and carboxyl groups participating in hydrogen bonds with the

water cluster. The water clusters in the N isomers contain many identifiable 5- and 6-member ring structures, suggesting that the water molecules are still preferentially hydrogen bonding with each other. All levels of theory predict Z24-1 to be the global minimum. The energy separation between Z24-1 and N24-1 is 4.8, 0.4, and 0.9 kcal/mol at the MP2//EFP2, MP2(EFP1/DFT) optimized, and PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p) levels of theory, respectively. This small separation may suggest that the true Z minimum has not been found for n = 24. The preferential stabilization of N24-1 relative to Z24-1 by MP2(EFP1/DFT) optimization occurs because MP2 causes tighter and stronger water-water binding. This is more important in the N structures than in the Z structures, given the strong ion-water interactions in the latter.

Alanine(H$_2$O)$_{32}$ Figure 14

The lowest energy 32-water N isomer, N32-1, does not appear to be fully solvated. However, with 32 waters in the system, it is possible to find structures, such as N32-4, that appear to be fully solvated; that is, completely surrounded by water molecules. At the highest level of theory, N32-4 is ~15 kcal/mol above N32-1, so the fully solvated species is clearly not the lowest energy N isomer. This may be due to the resistance of the hydrocarbon moiety to aqueous solvation. The two lowest energy Z isomers, Z32-1 and Z32-2 are separated by only 2 kcal/mol, and both are more fully solvated than the N species. The two lowest energy N isomers are 9.1 and 9.6 kcal/mol, respectively, higher in energy than Z32-1 at the highest level of theory. Boltzmann averaging of the alanine isomers at MP2(EFP1/DFT) and adding PCM yields a N-Z energy difference of 13.3 and 11.5 kcal/mol respectively. The experimental $\Delta H_{N \rightarrow Z}$ of glycine is 10.3

kcal/mol.[66] The $\Delta H_{N \rightarrow Z}$ of alanine is expected to be similar to this value. This suggests that the solvation of alanine is not converged at 32 waters and that addition of more waters is needed to converge the alanine $\Delta H_{N \rightarrow Z}$ for solvated N and Z forms.

Discussion

This study addresses the determination of the global minima for alanine as a function of the number of water molecules that are present in the system. The justification for using the EFP2 method to predict geometries is based upon favorable comparisons with MP2 calculations on alanine with 1- 3 waters. EFP2 reproduces the MP2 structures, as well as the relative energies of N isomers (and separately, Z isomers) among themselves. Given the semi-classical nature of the EFP method, one cannot expect this method to provide accurate N vs. Z relative energies. The agreement between EFP2 and MP2 optimizations generally continues for larger numbers of waters. However, there are cases for which the energy ordering changes upon MP2 optimization, because MP2 binds the waters more strongly to each other and to alanine. However, the identification of the Z and N minima is consistent with both EFP2 and MP2 optimizations, so employing the EFP2 method to determine preliminary geometries for low-lying structures is a good strategy.

Geometry optimization within a continuum solvent model has a negligible effect on the N alanine structure whereas there are significant effects on the Z structure, as has been reported previously.[17] Proper determination of Z structures in solution requires optimization in the presence of the solvent. This geometry effect decreases as the number of explicit waters increases. For example, with 32 waters present, the deviations from EFP2 structures are almost exclusively due to the changes in the water orientations. Optimiza-

tion via MP2(EFP1/DFT) with single point calculations using PCM at these geometries simulates the effect of adding a bulk solvent around the alanine(H_2O) complex.

Table 3 presents the results of a series of MP2 calculations that illustrate the importance of including dynamic correlation on the relative energies. The change between columns one and two derives from the treatment of the water as MP2. While this does change the Z-N relative energy, the qualitative trend is the same for both sets of calculations. In column three, the internal geometry of the alanine is allowed to relax in the optimization (recall that EFP internal structures are frozen). This relaxation does not alter the prediction of N or Z as the global minimum, but the Z-N relative energy no longer changes monotonically as the number of waters increases. In column four, the entire system is treated with MP2 at the MP2(EFP1/DFT) geometries. The relative energies are very similar to those in column two, and predicts a monotonic increase in the relative stability of Z vs. N as the number of water molecules increases. This suggests once again the importance of electron correlation in predicting the relative energies of these species.

Table 4 illustrates the changes in the energies relative to the global minimum as the level of theory is improved, for n = 1-32 water molecules. The HF//EFP2 level of theory is improved upon in several ways: by adding electron correlation, by re-optimizing the geometry with the alanine treated by MP2, and by adding the continuum solvent. HF//EFP2 predicts that the Z form remains higher in energy than the N form until 16 waters are present. The addition of electron correlation, with the alanine treated as DFT or MP2, predicts that seven waters are required for the Z form to become the global minimum. In all three of these cases, the addition of more water molecules systematically lowers the energy of the lowest-lying Z isomer relative to that of the lowest-lying N iso-

mer. The behavior of this N-Z relative energy is not as monotonic when the structures are re-optimized with MP2 (EFP1/DFT); however, the overall trend (increasing relative stability of Z as the number of water molecules increases) is retained. Addition of PCM (last two columns in Table 4) greatly increases the relative stability of Z alanine at smaller numbers (2-3) of waters.

Table 5 shows the change in the N-Z relative energy averaged over the energy difference between each N structure relative to each Z structure, due to the addition of correlation corrections and the continuum solvent, for the structures determined by EFP2 optimizations. For 1-8 waters, Aikens[17] found that electron correlation provided by MP2 stabilizes Z relative to N by 7-9 kcal/mol. As shown in Table 5, using EFP2 structures, the MP2 contribution to the N-Z energy difference decreases steadily with additional waters, from 6.8 kcal/mol for one water to 2.5 kcal/mol for eight waters to 0.9 kcal/mol for 32 waters. The second column in Table 5 illustrates that the large effect of the continuum on the N-Z energy difference diminishes to essentially zero by the time 32 explicit water molecules are present.

Binding energies

Boltzmann averaged energies were calculated for each water cluster using the Boltzmann equation:

$$\frac{\sum_i X_i e^{-\Delta E_i / RT}}{\sum_i e^{-\Delta E_i / RT}} = X_n$$

where X_i is the calculated property (e.g., the energy) of the ith structure, corrected with the zero point vibrational energy. ΔE_i is calculated by taking the difference between the energy of the ith structure and the energy of the lowest energy structure for a given n; T =

273 K. X_n is the Boltzmann averaged value of X for all structures for a given number (n) of waters. The total binding energies (shown in Table 6) may then be calculated:

$$B_{e(n)}= E[ALA(H_2O)_n]- n*E(H_2O) - E[ALA]$$

where ALA = Alanine and n = 0-32.

The differential binding energy is defined as the energy difference for the following process:

$$\Delta D_e = B_{e(n)} - B_{e(n-1)}$$

where n = 0-8. The differential binding energies (Table 7) were calculated by taking the Boltzmann averaged energy for n waters and subtracting this from the sum of the Boltzmann averaged energy for n-1 waters plus the energy of one water molecule. The change in binding energies is of interest to study the convergence of the solvation shell.

The Boltzmann averaged differential binding energies, the energy associated with systematically adding one water, were obtained using the EFP2 Hessians for all levels of theory. A validation of this approach was carried out for the lowest energy conformers (N and Z) for 3-5 waters, by performing MP2 optimizations and Hessian calculations for the entire alanine + water system. On average, the relative energies obtained after adding the EFP2 zero point energy (ZPE) correction to the EFP2 relative energies are in agreement with the corresponding MP2 values to within ~1 kcal/mol.

Total binding energies are shown in Table 6. The addition of correlation energy generally increases the total binding energies. Both DFT and MP2 are in very good agreement with each other.

The differential binding energies shown in Table 7 fluctuate as the number of water molecules is increased. The N differential binding energy has large fluctuations until

the addition of five waters. Subsequently, the fluctuations are small, ~1 kcal/mol. The fluctuations for the Z isomer do not significantly decrease by eight waters. This fluctuation in differential binding energies has also been observed, both experimentally and computationally, for the systematic addition of water molecules to small anions.[66] The DFT and MP2 differential binding energies are in good agreement with each other.

Conclusions and Summary

The generalized effective fragment potential (EFP2) captures reasonably well the geometries of the fully *ab initio* alanine-water complexes, so the EFP2 method provides an efficient method for obtaining these structures, especially as the number of water molecules is increased. The EFP2 Z vs. Z and N vs. N relative energies agree well with those obtained with MP2(EFP1/DFT)//EFP. The relative energies of Z isomers vs. N isomers are difficult for this method to capture, given the substantially different electron distributions of these two species.

Electron correlation is essential in order to correctly predict the relative energies of the lowest-energy N and Z isomers. Both DFT/B3LYP and MP2 provide substantial stabilization of Z alanine, especially for small numbers of explicit water molecules. Upon the addition of correlation corrections, the Z form becomes lower in energy at seven water molecules, for both DFT/B3LYP and MP2 single points at EFP2 geometries. When the continuum solvent is included in the calculations, the Z species become lower in energy with the addition of just 2-3 water molecules. As the number of discrete water molecules increases, the importance of the continuum solvent on the energy difference between the N and Z forms decreases to nearly zero. With 16 or more waters present, a large portion of the long range interactions with the bulk solvent are therefore being re-

covered.

The large number of low energy structures illustrates the importance of configurational sampling. So, many N and Z structures must be taken into account via statistical averaging if one is to calculate properties of even small clusters.

It is useful to compare the main findings presented here with those of previous researchers. Chuchev and BelBruno found (for clusters with 10 waters) a 6.6 and 5.2 kcal/mol heat of formation for the N $\rightarrow$ Z transition ($\Delta H_{N \rightarrow Z}$) using DFT/B3LYP and MP2 single point energies, respectively, at the DFT geometries.[39] Kwon, Kim and No[27] found $\Delta H_{N \rightarrow Z}$ of 25.0 and 25.6 kcal/mol for one and two waters, respectively, at the HF level of theory. Rzepa and Yi[14] found $\Delta H_{N \rightarrow Z}$ to be 9.1 and 5.6 kcal/mol for 7 and 15 waters, respectively, with PM3[83] calculations, and 14.8 and 1.0 kcal/mol for 7 and 15 waters, respectively, with the AM1 method.[84] These authors used a SCRF continuum model with the PM3 and AM1 semi-empirical methods to obtain $\Delta H_{N \rightarrow Z}$ = 7.4 and 2.9 kcal/mol, respectively. The global minimum MP2 [Boltzmann averaged MP2] $\Delta H_{N \rightarrow Z}$ predicted in this work for 32 waters (see Table 4) is 11.7 [12.6] kcal/mol, 10.1 [13.3] kcal/mol or 9.1 [11.5] kcal/mol using MP2(EFP1/DFT)//EFP2, MP2(EFP1/DFT) or PCM+MP2//MP2(EFP1/DFT), respectively. The experimental value of $\Delta H_{N \rightarrow Z}$ for glycine is 10.3 kcal/mol.[66] One would expect a similar value for alanine.

The first solvation shell first appears to begin forming in the Z form somewhere between 24 and 32 waters. When starting MC with a fully surrounded structure for 24 waters, the MC always found the waters clustered on the COOH/COO⁻ and NH$_2$/NH$_3^+$ moieties of the N and Z alanine. With 32 waters, there are N structures found by MC that exhibit a solvated structure, including solvation of the methyl group laying 20-30

kcal/mol above the N minimum. Therefore the $\Delta H_{N \rightarrow Z}$ energies presented may not quantitatively capture the $Z_{(aq)} \rightarrow N_{(aq)}$ process, since the lowest energy N structures are only partially solvated.

The goal of finding a converged N-Z energy difference is not complete. The N to Z energy difference appears to be still increasing with additional waters. However, the $\Delta H_{N \rightarrow Z}$ for 32 waters is within a few kcal/mol of the glycine experimental value for methods including electron correlation. It is possible that convergence of the energy difference will coincide with the emergence of a fully solvated N species as the N minimum. Furthermore, MD calculations on alanine would allow for the prediction of the properties (dipole moment, diffusion coefficient, density, heat of vaporization) seen in solvated species, which may be of interest to those developing models for amino acids.

Acknowledgments

References

1. Vishveshwara, S.; Pople, J. A. *J. Am. Chem. Soc.* **1977**, *99*, 2422.
2. Tse, Y-C.; Newton, M. D.; S. Vishveshwara, S.; Pople, J. A. *J. Am. Chem. Soc.* **1978**, *100*, 4329.
3. Clementi, E.; Cavallone, F.; Scordamaglia, R. *J. Am. Chem. Soc.* **1977**, *99* 5531.
4. Clementi, E.; Computational Aspects for large Chemical Systems, Lecture notes in Chemistry, Vol. 19, Springer, Berlin 1980.
5. Jensen, J.H.; Gordon, M. S. *J. Am. Chem. Soc.* **1991**, *113*, 7917.
6. Hu, C-H.; Shen, M.; Schaefer, H. F., III. *J. Am. Chem. Soc.* **1993**, *115*,106.
7. Ding, Y.; Krogh-Jespersen, K. *J. Comput. Chem.* **1996**, *17*, 338.
8. Gontrani, L.; Mennucci, B.; Tomasi, J. *J. Mol. Struct.: THEOCHEM.* **2000**, *500*, 113.
9. Bonaccorsi, R.; Palla, P.; Tomasi, J. *J. Am. Chem. Soc.* **1984** , *106*, 1945.
10. Truong, T. N.; Stefanovich, E. V. *J. Chem. Phys.* **1995**, *103*, 3709.
11. Andzelm, J.; Kolmel, C.; Klamt, A. *ibid.* **1995**, *103*, 9312. - later

12. Adamo, C.; Dillet, V.; Barone, V.; *Chem. Phys. Lett*. **1996**, *263*, 113.
13. Tortonda, F. R.; Pascual Ahuir, J-L.; Silla, E.; Tunon, I.; Ramirez, F. J. *J. Chem. Phys*. **2000**, *109*, 592.
14. Tunon, I.; Silla, E.; Ruiz-Lopez, M. F. *Chem. Phys. Lett*. **2000**, *321*, 433
15. Gontrani, L.; Mennucci, B.; Tomasi, J. *THEOCHEM* **500**, *113* 2000.
16. Bandyopadhyay, P.; Mennucci, B.; Tomasi, J.; Gordon, M.S. *J. Chem. Phys.;* **2002**, *116*, 5023.
17. Aikens, C. M.; Gordon, M. S. *J. Am. Chem. Soc*., **2006**, *128*, 12835.
18. Jensen, J. H.; Gordon, M. S. *J. Am. Chem. Soc*. **1995**, *117*, 8159.
19. Rzepa, H. S.; Yi, M. *J. Chem. Soc. Perkin Transactions*. **1991**, *2*, 531.
20. Cui, Q. *J. Chem. Phys*. **2002**, *117*, 4720.
21. Bandyopadhyay, P.; and Gordon, M. S. *J. Chem. Phys*. **2000**, *113*, 1104.
22. De Vries, A. H.; Van Duijnen, P. T.; Juffer, A. H.; Rullmann, J. A. C.; Dijkman, J. P.; Merenga, H.; Thole, B. T. *J. Comp. Chem*. **1995**, *16*, 37.
23. Russell, S. T.; Warshel, A. *J. Mol. Biol*. **1985**, *185*, 389.
24. Fernandez-Ramos, A.; Smedarchina, Z.; Siebrand, W.; Zgierski, M. Z. *J. Chem. Phys*. **2000**, *113*, 9714.
25. Ramaekers, R.; Pajak, J.; Lambie, B.; Maes, G. *J. Chem. Phys*. **2004**, *120*, 4182.
26. Yamabe, S.; Ono, N; Tsuchida, N. *J. Phys. Chem A*. **2003**, *107*, 7915.
27. Campo, M. G. *J. Chem. Phys*. **2006**, *125*, 114511.
28. Santosh, K.; Rai, A. K.; Rai S. B.; Rai D. K. ; Singh, A. N.; Singh, V. B. *J. mol. struct*. **2006**, *791*, 23-29.
29. Topol, A; Burt, S. K.; Russo, N.; Toscano, M. *J. Am Soc Mass Spectrom* **1999**, *10*, 318.
30. Powis, I.; Rennie, E. E.; Hergenhahn, U.; Kugeler, O.; Bussy-Socrate, R. *J. Phys. Chem. A*, **2003**, *107*, 25.
31. Osted, A.; Kongsted, J.; Christiansen,O. *J. Phys. Chem. A*, **2005**, *109*, 1430.
32. Kikuchi, O.; Watanabe, T.; Ogawa, Y.; Takase, H.; Takahashi, O. *J.phys.org.chem.* **1997**, *10*, 145.

33. Messer, B. M.; Cappa, C. D.; Smith, J. D.; Drisdell, W. S.; Schwartz, C. P.; Cohen, R. C.; Saykally, R. J. *J. Phys. Chem. B* **2005**, *109*, 21640.
34. Zhang, H.; Zhou, Z.; Shi, Y. *J. Phys. Chem. A* **2004**, *108*, 6735.
35. Kwon, O. Y.; Kim, S. Y.; No, K. T. *Bul. Korean Chem. Soc*. **1995**, *16*, 410.
36. Ahn, D.; Park, S.; Jeon, I.; Lee, M.; Kim, N.; Han, Y.; Lee, S. *J. Phys. Chem. B* **2003**, *1*, 14109.
37. Park, S.-W.; Ahn, D.-S.; Lee, S. *Chem. Phys. Lett*. **2003**, *371*, 74.
38. Xu, S.; Nilles, J.; Bowen, K. H. *J. chem. phys*. **2003**, *119*, 10696.
39. Chuchev, K.; BelBruno, J. J. *J.Mol. Struct.: THEOCHEM. 2008, 850, 111.*
40. Ellzy, M.; Jensen, J.O.; Hameka, H. F.; Kay, J. G. *Spectrochimica Acta Part A* **2003**, *59*, 2619.
41. Tajkhorshid, E.; Jalkansen, K. J.; Suhai, S. *J.Phys. Chem. B*. **1998**, *102*, 5899.
42. Frimand, K.; Bohr, H.; Jalkanen, K.J.; Suhai, S. *J. Chem. Phys*.; **2000**, *225*, 165.
43. Saqarik, K; Dokmaisrijam, S. *theochem*. **2005**, *718*, 31.
44. Dixit, S. B.; Bhasin, R.; Rajasekaran, E.; Jayaram. B. *J. Chem. Soc., Faraday T rans*. **1997**, *93*, 1105.
45. Kemp, D.; Gordon, M. S. *J. Phys. Chem. A* **2008**, *112*, 4885.

46. Blanco, S.; Lesarri, A.; Lopez, J. C.; Alonso, J. L. *J. Am. Chem. Soc.* **2004,** *126*, 11675.

47. Miertus, S.; Scrocco, E.; Tomasi, J. *J. chem. phys.* **1981**, *55*, 117.

48. Cances, E.; Mennucci, B.; Tomasi, J. *J. chem. phys.* **1997**, *107*, 30323041.

49. Li, H.; Jensen, J.H. *J. Compu. Chem.* **2004**, *25*, 144.

50. Osagner, L. *J. Am. Chem. Soc.* **1938**, *58*, 1486.

51. Cramer, C. J.; Truhlar, D. G. *Accts. Chem. Res.* **2008**, *41*, 760.

52. Klamt, A.; Shüürmann, G. *J. Chem Soc. Perkin Trans.* **1993**, *II*, 799.

53. Baldridge, K.; Klamt, A. *J.Chem.Phys.* **1997**, *106*, 6622.

54. Zhan, C-G.; Bentley, J.; Chipman, D. M. *J. Chem. Phys.* **1998**, *108*, 177

55. (a)Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W.J. *J. Phys Chem A*, **2001**, *105,* 293. (b)

56. Jensen, J.H.; Gordon, M. S. *J. Chem. Phys.* **1996**, *104*, 7795.

57. Jensen, J. H.; Gordon, M. S. *Mol. Phys.* **1996**, *89*, 1313.

58. Jensen, J. H.; Gordon, M. S. *J. Chem. Phys.* **1998**, *108*, 4772.

59. Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, *105*, 11081.

60. (a) Day, P.N.; Jensen, J.H.; Gordon, M.S.; Webb, S.P.; Stevens, W.J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968. (b) Stone, A.J. *The Theory of Intermolecular Forces*; Oxford Press; 1996.

61. Adamovic, I.; Freitag M. A.; Gordon, M. S. *J. Chem. Phys.* **2003**,*118*, 6725.

62. Song, J; Gordon, M. S. unpublished data.

63. Gordon, M.S.; Slipchenko, L.; Li, H.; Jensen, J. H. *Ann. Rep. Comp. Chem.* **2007**, *3*, 177.

64. Adamovic, I.; Gordon, M. S.; *Mol. Phys.* **2005**, *103*, 379.

65. Li. H.; Jensen, J. H.; Gordon, M. S.; *J. Chem. Phys,* **2006**, *124*, 214108.

66. Wada, G.; Tamura, E.; Okina, M.; Nakamura, M. *Bul. Chem. Soc. Jpn.* **1982**, *55*, 3064

67. (a)Shoemaker, D.P.; Donahue, J.; Schoemaker, V.; Corey, R.B. *J. Am. Chem. Soc.* **1950**, *72*, 2328. (b) Lehman, M. S.; Kotzle, T. F.; Hamilton, W. C. *J. Am. Chem. Soc.* **1972**, *94*, 2657.

68. (a)Kvik, A.; Canning, W. M.; Koetzle, T. F.; Williams, J. B. *Acta Crystallogr.* **1980** *B36*, 115. (b) Destro, R.; Marsh, R. E.; Bianchi, R. *J. Phys. Chem.* **1988**, *92*, 854.

69. Jiang, J-S.; Brünger, A. T. *J. Mol. Biol.* **1994**, *243*, 100

70. Rupley, J.A.; Timasheff, S.N.; Fasman, G.D.; (Eds.), Structure and Stability of Biological Macromolecules, Marker Dekker, New York, 1969, pp. 291-352

71. Mathews, B.W.; Neurath, H.; Hill, R.L.; (Eds.) The Proteins vol. III, Academic Press, New York, 1977, pp. 403-590.

72. Roothan, C. C. *J. Rev. Mod. Phys.* **1960**, *32*, 179

73. Moller, C.; Plesset, S. *Phys. Rev.* **1934**, *46*, 618

74. Hehre, W. J.; Ditchfield, R.; Pople, J.A. *J. Chem. Phys.* **1972**, *56*, 2257

75. Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *J. Chem. Phys.* **1982**, *77*, 3654.

76. Hariharan, P. C.; Pople, J. A. *Theoret. Chim. Acta* **1973**, *28*, 213.

77. Metropolis, N.; Rosenbluth, A.; Teller, A. *J. Chem. Phys.* **1953**, *21*, 1087. (b) Kirkpatrick, S.; Gelatt, C. D.; Vecci, M. P. *Science.* **1983**, *220*, 671 (c) Wales, D.

J.; Hodges, M. P. *Chem.Phys.Lett*. **1998**, *286*, 65. (d) Li, Z.; Scheraga, H. A. *Proc.Nat.Acad.Sci*. **1987**, *84*, 6611. (e) Parks, G. T. *Nucl. Technol*. **1990**, *89*, 233.

78. Day, P.N.; Pachter, R.; Gordon, M. S.; Merrill, G. N. *J. Chem. Phys*. **2000**, *112*, 2063.

79. Becke, A. D. *J. Chem. Phys*. **1993**, *98*, 5642 (b) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 78.

80. . (a) 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.; *J. Comput. Chem.* **1993**, *14*, 1347.;(b) Gordon, M.S.; Schmidt, M.W., *Theory and Applications of Computational Chemistry, the first forty years*; Elsevier; Amsterdam; 2005

81. Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Mod*. **1998**, *16*, 133.

82. Bachrach, S. M. *J. Phys. Chem. A* **2008**, *112*, 3722.

83. Stewart, J. J. P. *J. Comput. Chem*. **1989**, *10*, 209. (b) Stewart, J. J. P. *J. Comput. Chem*. **1989**, *10*, 221. (c) Stewart, J. J. P. *J. Comput. Chem*. **1991**, *12*, 320.

84. Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. *J. Amer. Chem. Soc*. **1985**, *107*, 3902.

Figure Captions

Figure 1. L form neutral alanine conformer structures and energies (relative to ALA 1 L) at RHF/6-31++G(d,p), MP2/6-31++G(d,p), PCM+RHF/6-31++G(d,p), PCM+RHF/6-31++G(d,p) levels of theory.

Figure 2. D form neutral alanine conformer structures and energies (relative to ALA 1 D) at RHF/6-31++G(d,p), MP2/6-31++G(d,p), PCM+RHF/6-31++G(d,p), PCM+MP2/6-31++G(d,p) levels of theory.

Figure 3. Alanine zwitterion and the three rotamers (45˚, 90˚ rotation of the COO$^-$ group about the C-C bond, 60˚ rotation of NH$_3^+$ group about the N-C bond) used in Monte Carlo simulations.

Figure 4. Alanine and one water. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the N form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 5. Alanine and two waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the N form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 6. Alanine and three waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global

minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 7. Alanine and four waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 8. Alanine and five waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 9. Alanine and six waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 10. Alanine and seven waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 11. Alanine and eight waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 12. Alanine and 16 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 13. Alanine and 24 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 14. Alanine and 32 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Table 1. Isomers of L and D neutral alanine, and the zwitterionic conformer. Effects of correlation and continuum solvation are added via MP2 and PCM respectively. The zwitterions are only stable when using PCM so no results are reported for HF or MP2. ZW' (90˙ COO⁻), ZW'' (45˙ COO⁻), ZW''' (60˙ NH$_3^+$) are rotamers and the values are from a single point and not optimized. Energies are in kcal/mol.

	HF*	MP2*	RHF+PCM*	MP2+PCM*
ALA 1 L	0.0	0.0	0.0	0.0
ALA 1 D	0.0	0.0	0.0	0.0
ALA 2 L	-2.8	-0.7	0.4	-0.7
ALA 2 D	-2.8	-0.7	0.4	-0.7
ALA 3 L	-1.2	0.6	1.3	0.4
ALA 3 D	-1.2	0.6	1.3	0.4
ALA 4 L	-1.5	0.3	1.3	2.4
ALA 4 D	-1.5	0.3	1.3	2.4
ALA 5 L	-0.5	0.3	1.9	2.3
ALA 5 D	-0.5	0.3	1.8	2.3
ALA 6 L	4.2	5.6	3.2	4.1
ALA 6 D	4.2	5.6	3.1	4.1
ALA 7 L	5.4	6.4	3.8	4.5

ALA 7 D	5.4	6.4	3.7	4.1
ALA 8 L	*5.8*	*6.8*	*4.1*	*4.9*
ALA 8 D	6.0	7.0	4.4	5.1
ZW			0.0	.0.0
ZW'			5.2	4.6
ZW"			5.4	5.1
ZW'''			2.0	2.6

*6-31++G(d,p)

Table 2. Global minimum for D alanine subtracted from the L alanine global minimum.

Energies are in kcal/mol.

$(H_2O)_n$	EFP2*	MP2(EFP 1/DFT)// EFP2*	MP2// EFP2*	MP2// MP2(EFP 1/DFT)*
3	2.9	-0.8	0.5	-0.1
4	0.0	0.9	1.8	0.8
5	0.3	2.2	3.0	1.0
6	5.2	3.1	2.1	1.3
7	0.9	0.4	-0.1	1.2
8	0.5	-0.8	0.9	1.6
16	1.7	1.3		
24	1.7	2.1		
32	0.1	1.6		

*6-31++G(d,p)

Table 3. Impact of electron correlation on the energies (kcal/mol) relative to the global minimum for n=1 - 8. In the first column alanine is MP2 and water is EFP1/DFT at the EFP2 geometry. The second column uses the EFP2 geometry but with both the water and alanine calculated via MP2. The third column represents an optimization in which the alanine is treated as MP2 and the water as EFP1/DFT. The fourth column is a single point at the MP2(EFP1/DFT) optimized geometries but with both the water and alanine calculated via MP2. All energies are relative to the global minimum for each water n=1 - 8 at the respective level of theory.

	MP2(EFP1/DFT) //EFP2*	MP2//EFP2*	MP2(EFP1/DFT)*	MP2[†]
$(H_2O)_n$	Z-N	Z-N	Z-N	Z-N
1	20.1	15.2	3.6	13.3
2	13.7	10.9	8.8	9.9
3	9.8	8.6	7.0	5.0
4	5.7	4.6	1.5	4.9
5	1.5	3.2	1.0	3.2
6	0.5	1.7	1.3	0.0
7	-0.2	-2.3	-2.0	-3.9
8	-0.8	-5.4	-2.1	-4.2

[*] 6-31++G(d,p) ZPE Corrected

[†] //MP2(EFP1/DFT)/6-31++G(d,p) ZPE Corrected

Table 4. Improvements to the HF energies (kcal/mol) relative to the global minimum for n=1-32. The first three columns are single point energies at the EFP2 optimized geometries. In the first column alanine is HF and water is EFP1/HF. In the second (third) columns alanine is DFT (MP2), and water is EFP1/DFT. The fourth column presents optimized geometries with MP2/6-31++G(d,p) alanine and EFP1/DFT water. Columns five and six use the MP2(EFP1/DFT) geometries and add the continuum solvent model (PCM). In both columns water is EFP1/DFT.

	HF(EFP1/HF) //EFP2*	DFT/B3LYP (EFP1/DFT)// EFP2*	MP2(EFP1/DFT) //EFP2*	MP2(EFP1/DFT)*	HF+PCM[†]	MP2+PCM[†]
$(H_2O)_n$	Z-N	Z-N	Z-N	Z-N	Z-N	Z-N
1	22.4	20.8	20.1	3.6	1.1	2.1
2	19.2	14.5	13.7	8.8	-0.9	1.8
3	14.7	6.3	9.8	7.0	-2.1	-1.3
4	11.9	5.1	5.7	1.5	-3.7	-1.8
5	10.2	2.0	1.5	1.0	-4.5	-4.1
6	8.5	1.7	0.5	1.3	-4.0	-2.2
7	5.5	-1.0	-0.2	-2.0	-5.4	-2.0
8	1.2	-1.1	-0.8	-2.1	-6.5	-2.5
16	-2.4	-2.0	-3.8	-15.6	-4.7	-5.2
24	-4.7	-4.7	-4.8	-0.4	-6.0	-0.9
32	-9.5	-10.2	-11.7	-10.1	-13.3	-9.1

* 6-31++G(d,p) ZPE Corrected

[†] //MP2(EFP1/DFT) 6-31++G(d,p) ZPE Corrected

Table 5. Average energy contributions to the N-Z energy difference for each N isomer relative to each Z isomer, for the addition electronic correlation energy (MP2) and continuum solvation (PCM) to the HF(EFP1/HF)//EFP2 N-Z relative energies.

$(H_2O)_n$	MP2 Correction	MP2+PCM Correction
1	6.8	29.9
2	6.7	25.7
3	5.3	20.1
4	4.6	16.6
5	3.9	12.9
6	3.1	9.6
7	4.3	9.6
8	2.5	9.3
16	1.3	2.7
24	0.3	0.8
32	0.9	0.3

Table 6. Total binding energy (kcal/mol) for N and Z alanine as a function of n waters at the HF, DFT/B3LYP, and MP2 levels of theory.

	Total Binding Energy - N				Total Binding Energy - Z		
n	HF	DFT	MP2		HF	DFT	MP2
1	-6.3	-8.5	-8.4		-13.0	-14.9	-15.1
2	-13.9	-17.9	-17.8		-23.7	-27.0	-27.9
3	-19.6	-24.7	-25.5		-33.5	-47.6	-45.7
4	-32.4	-38.7	-38.4		-43.1	-59.8	-60.7
5	-33.1	-43.0	-43.1		-51.8	-68.2	-68.7
6	-38.8	-51.9	-51.7		-67.9	-80.8	-81.3
7	-44.2	-64.9	-65.2		-71.8	-89.8	-90.5
8	-49.5	-67.9	-74.2		-81.2	-93.7	-98.3
16	-102.2	-143.4	-145.2		-135.3	-169.4	-170.4
24	-152.6	-201.9	-205.3		-190.4	-247.8	-248.8
32	-196.4	-272.4	-272.3		-225.4	-311.1	-312.2

Table 7. Differential binding energy (kcal/mol) for N and Z alanine as a function of n waters at the EFP2, HF, DFT/B3LYP, and MP2 levels of theory.

$(H_2O)_n$	(ΔDe) - Neutral					(ΔDe) - Zwitterion			
	EFP2	HF	DFT	MP2		EFP2	HF	DFT	MP2
1	-9.1	-7.6	-9.4	-9.4		-19.2	-10.7	-12.2	-12.7
2	-17.3	-5.4	-6.3	-7.4		-19.7	-9.8	-12.3	-10.0
3	-15.1	-12.9	-13.9	-12.8		-17.9	-9.6	-12.2	-14.8
4	-10.2	-0.7	-4.4	-4.8		-12.6	-8.7	-8.4	-8.0
5	-13.2	-5.8	-8.8	-8.5		-15.7	-7.4	-12.5	-12.7
6	-7.7	-5.3	-7.5	-7.5		-13.1	-8.6	-9.0	-9.2
7	-9.1	-5.4	-8.7	-8.5		-14.0	-9.4	-12.2	-12.3

Figure 1. L form neutral alanine conformer structures and energies(relative to ALA 1 L) at RHF/6-31++G(d,p), MP2/6-31++G(d,p), PCM+RHF/6-31++G(d,p), PCM+RHF/6-31++G(d,p) levels of theory.

ALA7L	5.4	6.4	3.8	4.5	ALA8L	5.8	6.8	4.1	4.9
ALA5L	-0.5	0.3	1.9	2.3	ALA6L	4.2	5.6	3.2	4.1
ALA3L	-1.2	0.6	1.3	.04	ALA4L	5.8	6.8	1.3	2.4
ALA1L	0.0	0.0	0.0	0.0	ALA2L	-2.8	-0.7	0.4	-0.7

Figure 2. D form neutral alanine conformer structures and energies (relative to ALA 1 D) at RHF/6-31++G(d,p), MP2/6-31++G(d,p), PCM+RHF/6-31++G(d,p), PCM+MP2/6-31++G(d,p) levels of theory.

ALA7D   5.4   6.4   3.8   4.1   ALA8D   6.0   7.0   4.1   5.1

ALA5D   -0.5   0.3   1.8   2.3   ALA6D   4.2   5.6   3.1   4.1

ALA3D   -1.2   0.6   1.3   .04   ALA4D   5.8   6.8   1.3   2.4

ALA1D   0.0   0.0   0.0   0.0   ALA2D   -2.8   -0.7   0.4   -0.7

Figure 3. Alanine zwitterion and the three rotamers (45°, 90° rotation of the COO⁻ group about the C-C bond, 60° rotation of NH_3^+ group about the N-C bond) used in Monte Carlo simulations.

ZW (L)

ZW (D)

ZW" (L) (45° COO⁻)

ZW' (L) (90° COO⁻)

ZW''' (L) (120° NH_3^+)

Figure 4. Alanine and one water. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the N form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z1-1

20.4 17.5 16.3 8.2 2.1

N1-1

0.0 0.0 0.0 0.0 0.0

N1-2

-6.5 -3.3 -3.9 -5.7 0.5

Z1-1 (alternative view)

N1-3

-2.2 0.4 -0.1 -0.1 2.7

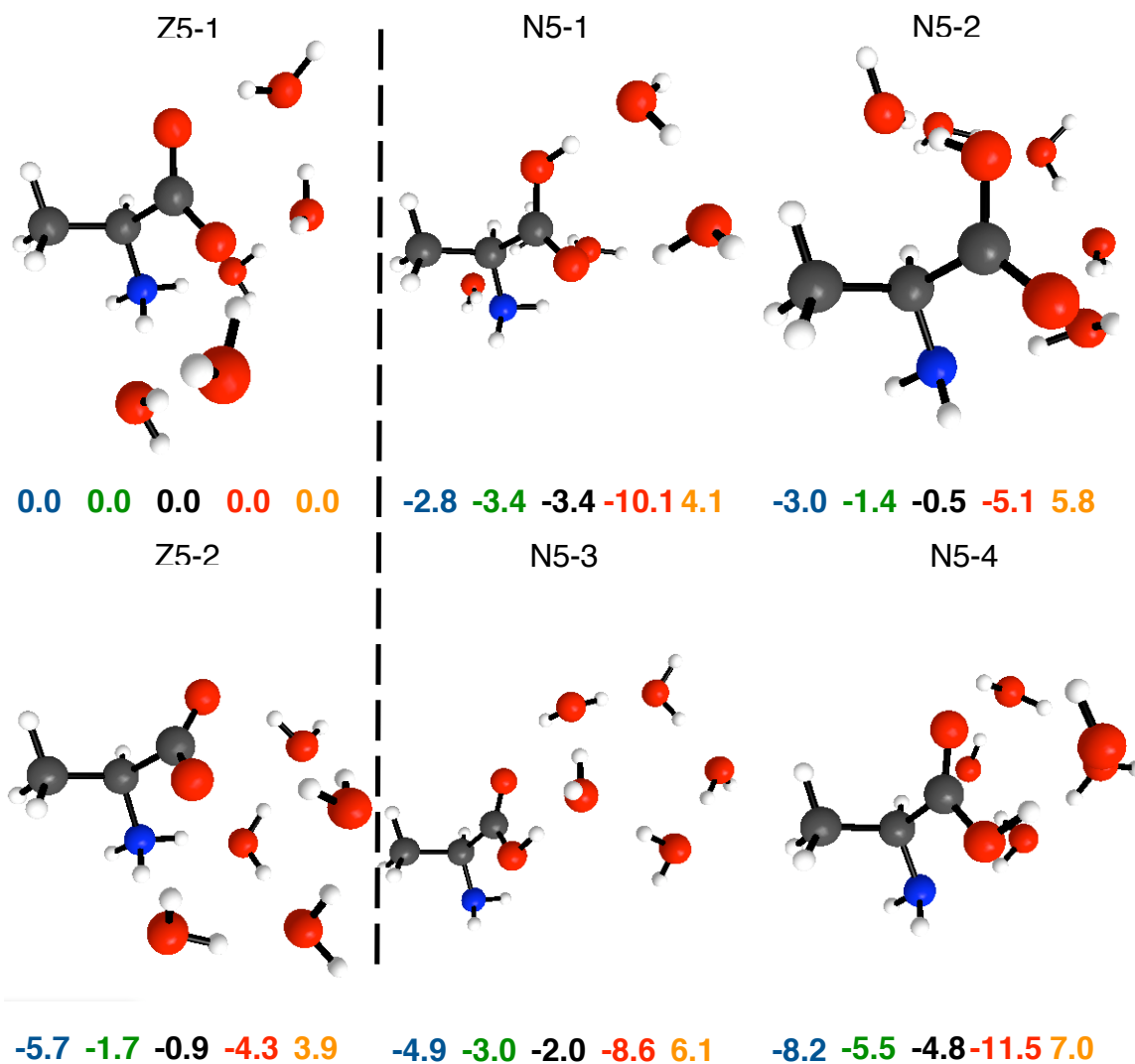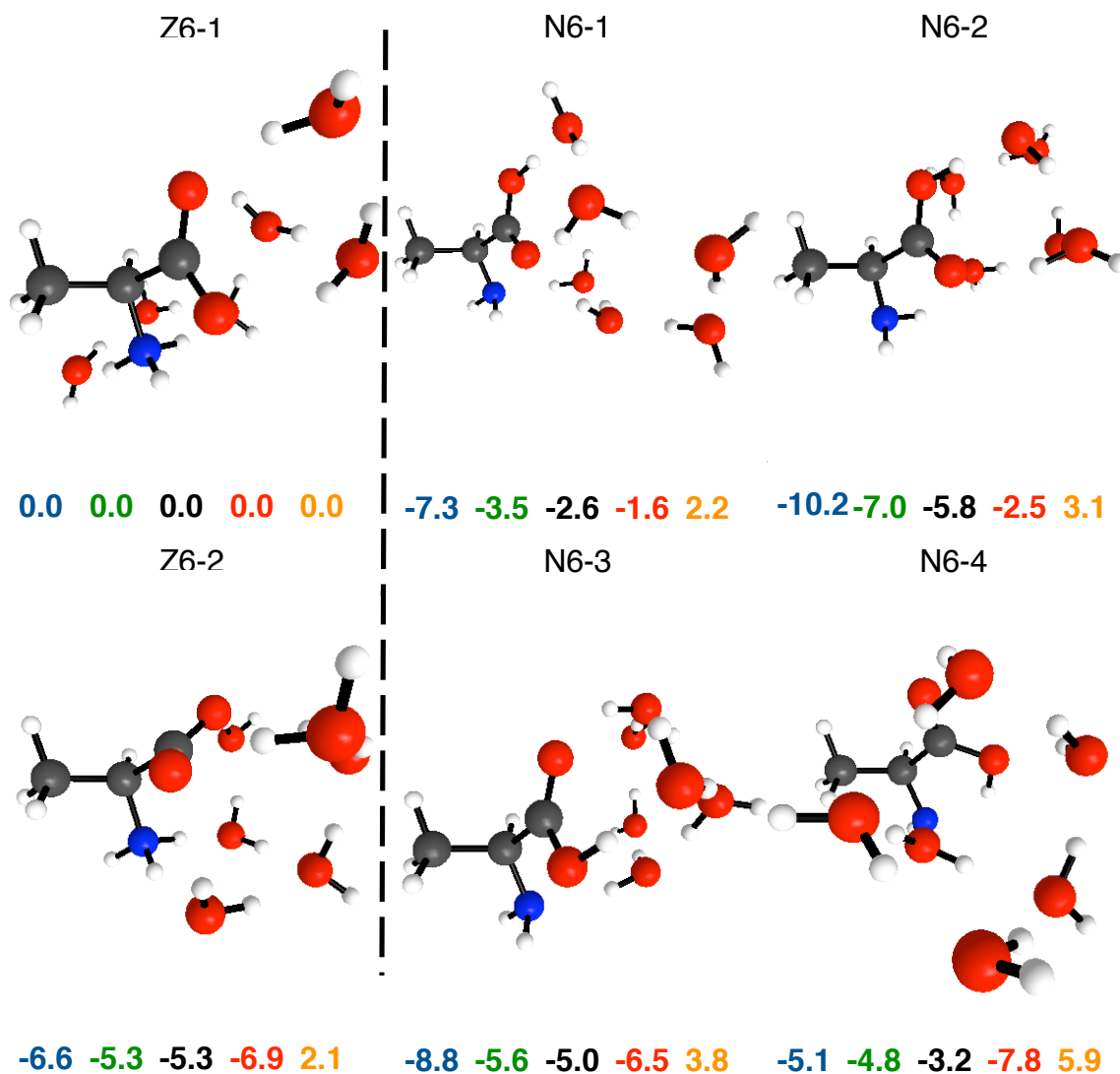HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 5. Alanine and two waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the N form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z2-1   N2-1   N2-2

20.6 16.4 15.3 13.9 1.8   **0.0 0.0 0.0 0.0 0.0**   **5.3 5.1 5.4 4.8 4.0**

Z2-2   N2-3   N2-4

20.5 14.5 13.7 10.2 1.8   **4.5 5.7 4.8 5.5 0.7**   **6.0 2.9 3.2 2.6 4.8**
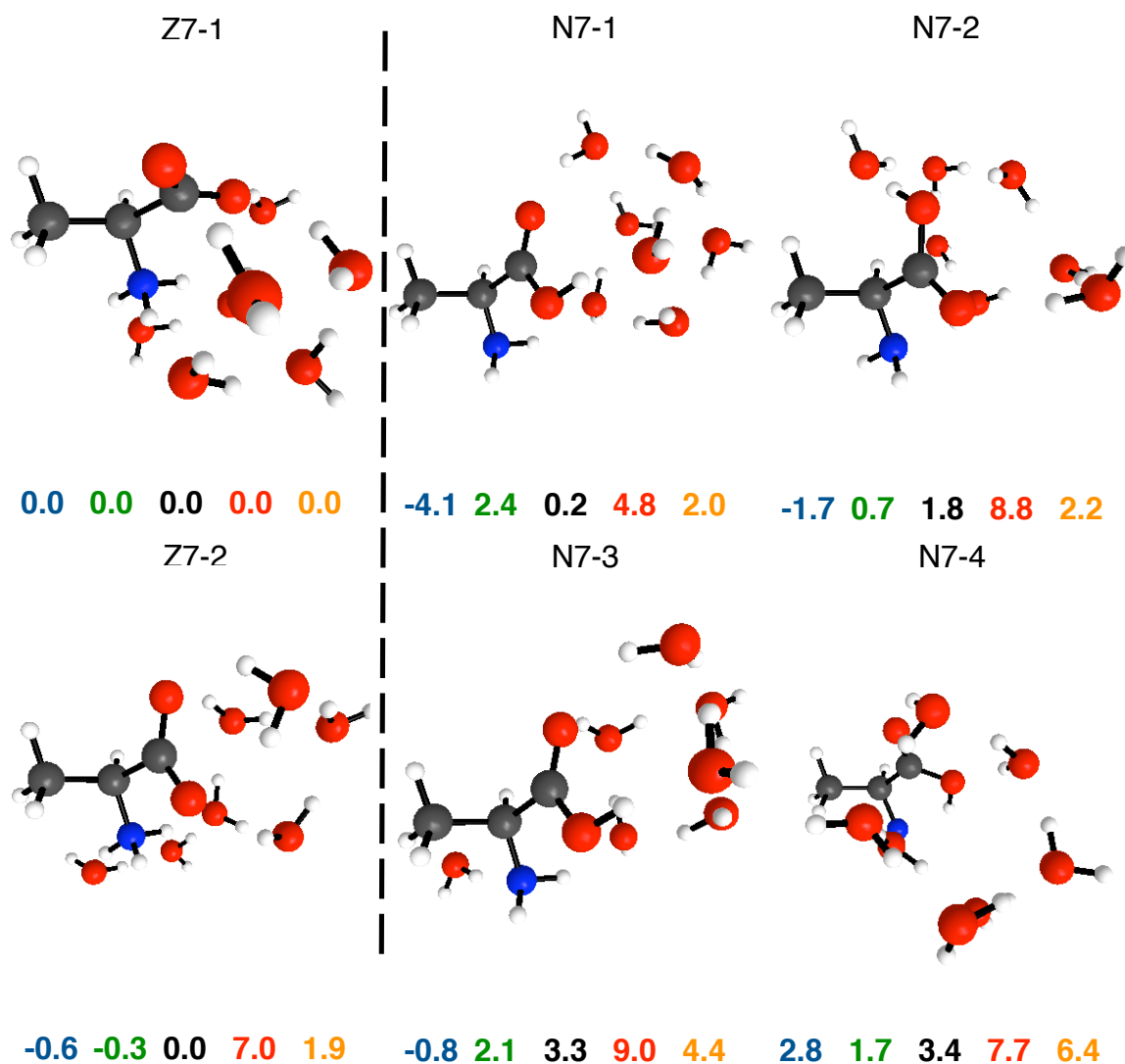
HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 6. Alanine and three waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z3-1   N3-1   N3-2

0.0 0.0 0.0 0.0 0.0   **-13.8 -8.3 -7.7 -7.7 1.3**   **-11.0 -7.6 -7.0 -7.4 1.4**

Z3-2   N3-3   N3-4

-0.7 -4.9 -0.7 -7.4 2.9   **-15.7 -11.2 -10.4 -11.2 1.9**   **-13.1 -8.6 -7.9 -8.7 2.4**

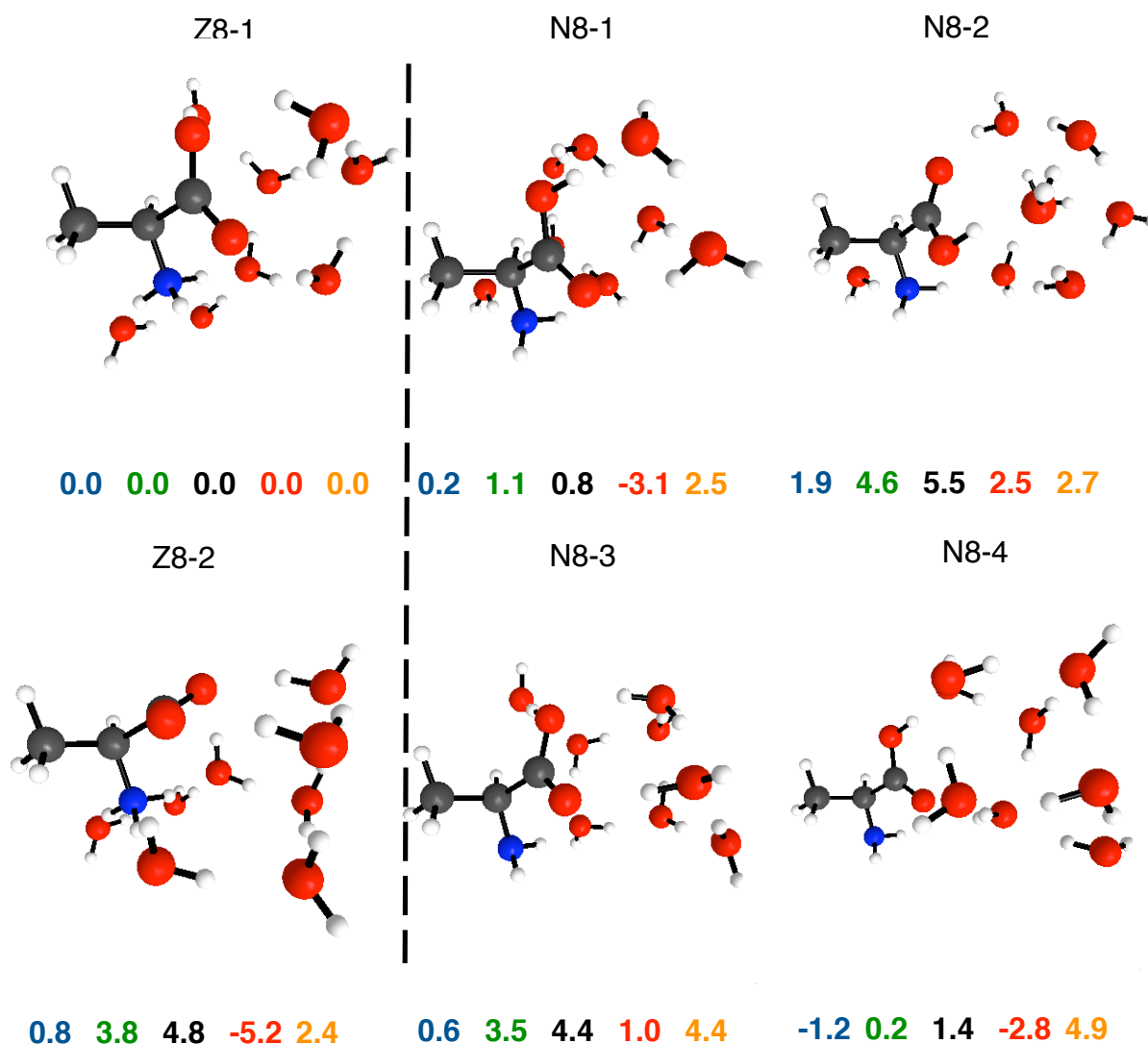HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 7. Alanine and four waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z4-1

0.0 0.0 0.0 0.0 0.0

N4-1

-10.5 -6.9 -5.8 -7.2 1.8

N4-2

-10.3 -6.4 -6.0 -4.2 4.9

Z4-2

-3.2 -3.3 -3.0 0.2 1.7

N4-3

-11.6 -6.9 -6.2 -6.2 8.7

N4-4

-13.5 -18.3 -8.7 -11.6 10.5
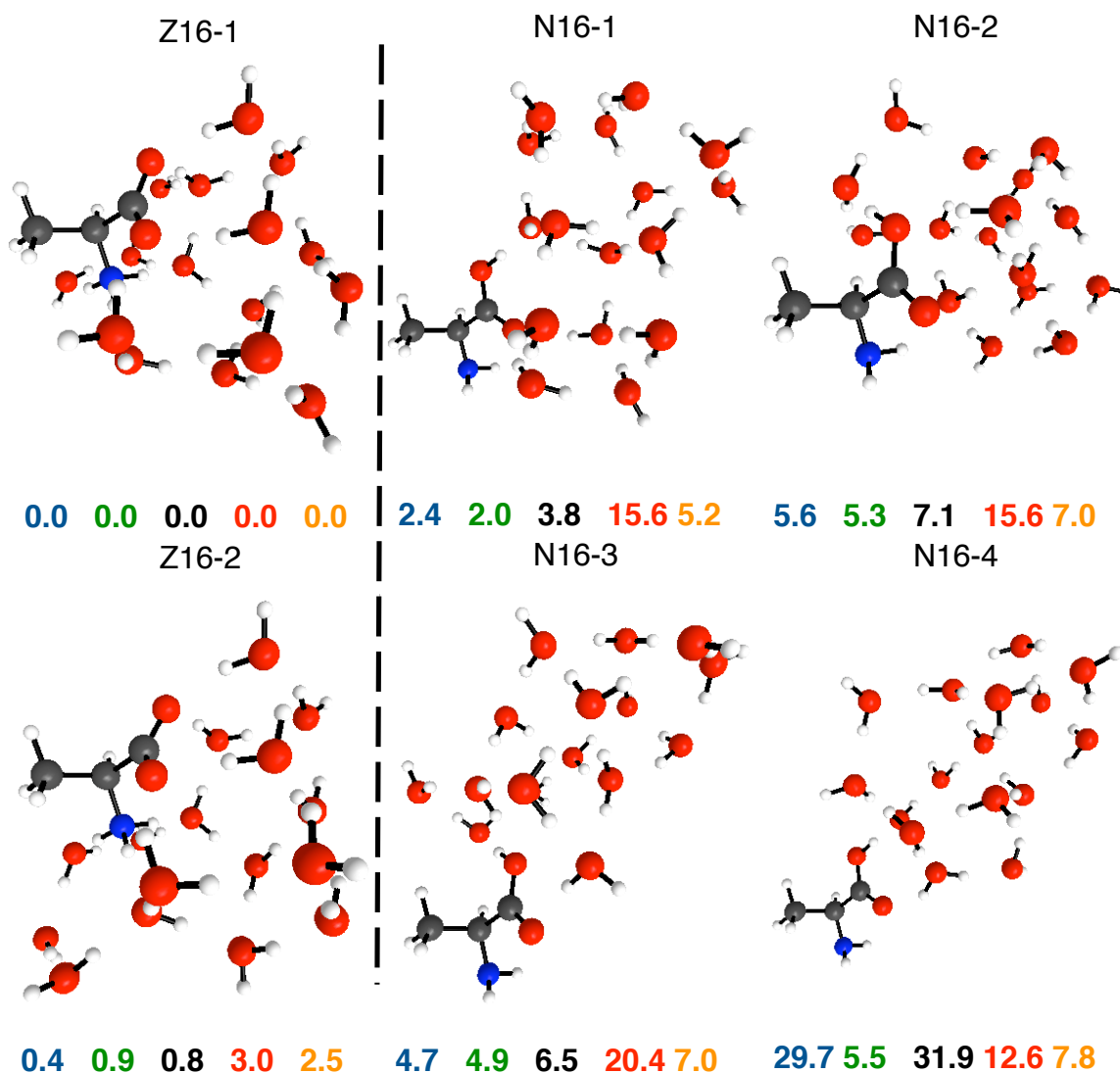
HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 8. Alanine and five waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z5-1

0.0 0.0 0.0 0.0 0.0

N5-1

-2.8 -3.4 -3.4 -10.1 4.1

N5-2

-3.0 -1.4 -0.5 -5.1 5.8

Z5-2

-5.7 -1.7 -0.9 -4.3 3.9

N5-3

-4.9 -3.0 -2.0 -8.6 6.1

N5-4

-8.2 -5.5 -4.8 -11.5 7.0

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
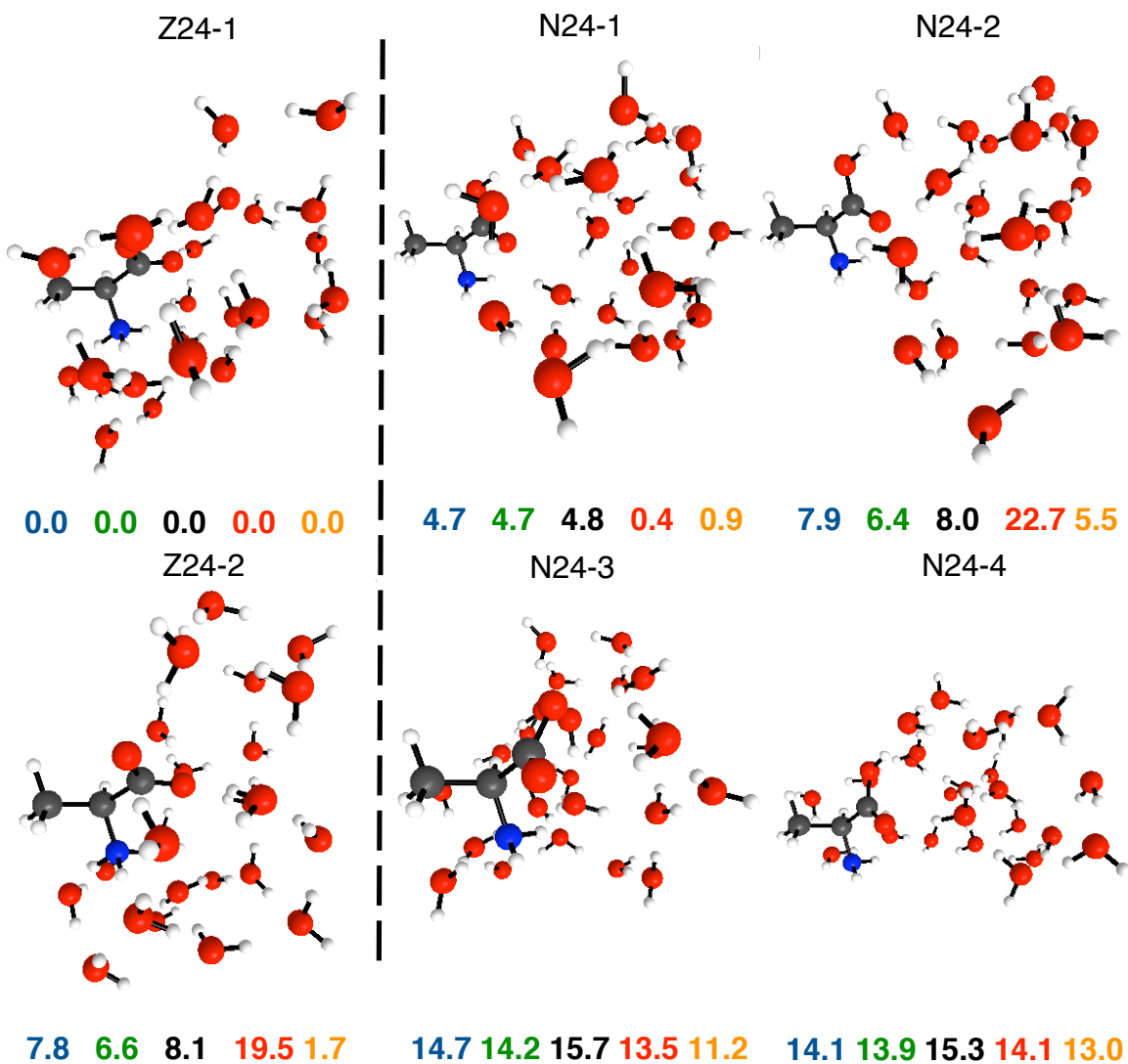MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 9. Alanine and six waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z6-1

N6-1

N6-2

0.0 0.0 0.0 0.0 0.0   -7.3 -3.5 -2.6 -1.6 2.2   -10.2 -7.0 -5.8 -2.5 3.1

Z6-2

N6-3

N6-4

-6.6 -5.3 -5.3 -6.9 2.1   -8.8 -5.6 -5.0 -6.5 3.8   -5.1 -4.8 -3.2 -7.8 5.9

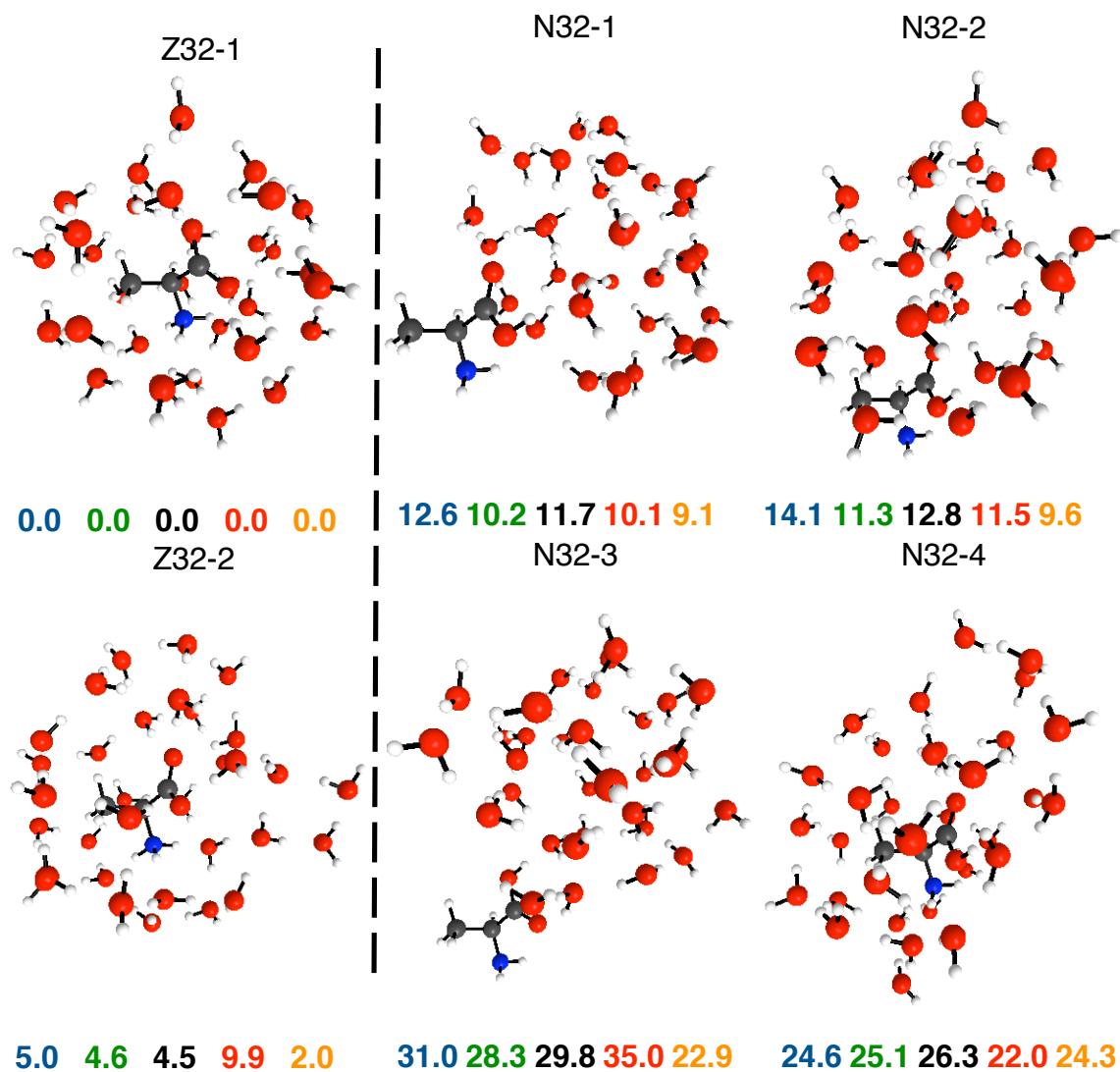HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 10. Alanine and seven waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

	Z7-1					N7-1					N7-2			
0.0	0.0	0.0	0.0	0.0	-4.1	2.4	0.2	4.8	2.0	-1.7	0.7	1.8	8.8	2.2

	Z7-2					N7-3					N7-4			
-0.6	-0.3	0.0	7.0	1.9	-0.8	2.1	3.3	9.0	4.4	2.8	1.7	3.4	7.7	6.4

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 11 Alanine and eight waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z8-1

0.0 0.0 0.0 0.0 0.0

N8-1

0.2 1.1 0.8 -3.1 2.5

N8-2

1.9 4.6 5.5 2.5 2.7

Z8-2

0.8 3.8 4.8 -5.2 2.4

N8-3

0.6 3.5 4.4 1.0 4.4

N8-4

-1.2 0.2 1.4 -2.8 4.9

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 12. Alanine and 16 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)/6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water)/6-31++G(d,p).

Z16-1   N16-1   N16-2

0.0 0.0 0.0 0.0 0.0   2.4 2.0 3.8 15.6 5.2   5.6 5.3 7.1 15.6 7.0

Z16-2   N16-3   N16-4

0.4 0.9 0.8 3.0 2.5   4.7 4.9 6.5 20.4 7.0   29.7 5.5 31.9 12.6 7.8

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 13. Alanine and 24 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z24-1

N24-1

N24-2

0.0 0.0 0.0 0.0 0.0

4.7 4.7 4.8 0.4 0.9

7.9 6.4 8.0 22.7 5.5

Z24-2

N24-3

N24-4

7.8 6.6 8.1 19.5 1.7

14.7 14.2 15.7 13.5 11.2

14.1 13.9 15.3 14.1 13.0

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 14. Alanine and 32 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N form minima with representative higher energy structures. Energies (kcal/mol) are relative to the Z form global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z32-1

0.0 0.0 0.0 0.0 0.0

N32-1

12.6 10.2 11.7 10.1 9.1

N32-2

14.1 11.3 12.8 11.5 9.6

Z32-2

5.0 4.6 4.5 9.9 2.0

N32-3

31.0 28.3 29.8 35.0 22.9

N32-4

24.6 25.1 26.3 22.0 24.3

HF (EFP1/HF water)//EFP2 6-31++G(d,p)
DFT (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2 (EFP1/DFT water)//EFP2 6-31++G(d,p)
MP2(EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

WATER AND ALANINE: FROM PUDDLES TO PONDS

Jonathan M. Mullin, Mark S. Gordon

Iowa State University, Ames Laboratory, Ames IA 50010

Abstract

The solvation of alanine is investigated, with a focus on adding a sufficient number of discrete water molecules to determine the first solvation shell for both the nonionized (N) and zwitterionic (Z) forms in order to converge the enthalpy of solvation and the enthalpy difference for the two forms of alanine. Monte Carlo sampling was employed using the generalized effective fragment potential (EFP) method to determine the global minimum of both conformers, with the number of EFP water molecules ranging from 32-49. A subset of sampled geometries were optimized with second order perturbation theory (MP2) using the 6-31++G(d,p) basis set. Single point energies were calculated at these geometries using the polarizable continuum model (PCM). The predicted 298.15K enthalpy of solvation for MP2/6-31++G(d,p) and MP2+PCM//MP2/6-31++G(d,p) ranges from 10.0-13.2 kcal/mol and 10.1-12.6 kcal/mol, respectively.

Introduction

Amino acid solvation is of considerable interest due to its role in biology and biochemistry. Neutral amino acids exist predominantly in their nonionized form (N) in the gas phase, while in solution or in crystals alanine is observed in the zwitterionic form (Z).[1] Varying the pH can cause alanine to exist as N, Z, anionic, or cationic species. The transition from neutral to zwitterion has been shown to be a consequence of the greater

stabilization of the Z relative to the N species, due to the formation of hydrogen bond-ing networks. The experimental free energy and enthalpy for the process $Z_{(aq)} \rightarrow N_{(aq)}$ are 7.3 and 10.3 kcal/mol, respectively, for glycine,[2] while the corresponding experimental values for alanine do not appear to be available.

The N $\rightarrow$ Z transition has been studied by systematic addition of water mole-cules for glycine and alanine previously.[3-5] Aikens and Gordon used Møller-Plesset per-turbation theory (MP2)[6] single point energies at restricted Hartree Fock (RHF)[7] geo-metries with a 631++G(d,p)[8] basis set, MP2//RHF/631++G(d,p). The polarizable contin-uum model (PCM)[9] was included to model bulk solvation. These authors determined 1.8 and 6.0 kcal/mol enthalpies for the process N $\rightarrow$ Z ($\Delta H_{N \rightarrow Z}$) using MP2//RHF/631++G(d,p), and MP2+PCM//RHF/631++G(d,p) respectively with eight wa-ter molecules[3]. Aikens and Gordon noted that convergence had probably not been achieved at this number of water molecules. Chuchev and BelBruno found a 6.6 and 5.2 kcal/mol enthalpy for the alanine N $\rightarrow$ Z transition ($\Delta H_{N \rightarrow Z}$) using density functional the-ory (DFT/B3LYP)[10] and MP2 single points, respectively, at the DFT geometry for 10 wa-ters.[4] The authors of the present work studied alanine in the presence of up to 32 effective fragment potential (EFP1)/DFT[11] waters and found $\Delta H_{N \rightarrow Z}$ = 11.7 kcal/mol, 10.1 kcal/mol or 9.1 kcal/mol using MP2//EFP2, MP2(EFP1/DFT) or PCM+MP2//MP2(EFP1/DFT), respectively, with 32 waters.[5]

The aforementioned studies used discrete explicit solvent models in order to in-vestigate the hydrogen bonded interaction between water molecules and the amino acid, adding a continuum (implicit) solvent as a third layer to simulate complete solvation. This work builds upon the previous analysis that included up to 32 discrete water mole-

cules, and additionally explores the importance of including a continuum description of the solvent. The salient aspects of the 32-water calculations are included here as a starting point for further investigation of up to 49 explicit water molecules.

Aikens and Gordon[3], Chuchev and BelBruno[4] and the authors[5] identified the Z species as the global minimum structure with the addition of seven waters when dynamic correlation (e.g., DFT or MP2) is included. Neither Aikens and Gordon nor Chuchev and BelBruno obtained a converged $\Delta H_{N \to Z}$ with discrete waters, although the latter authors, using MP2+PCM, observed a $\Delta H_{N \to Z}$ that oscillated between 6.0 and 8.8 kcal/mol for addition of 3-8 waters. In the previous systematic study of up to 32 waters[5], it was found that the alanine Z global minimum appears to be fully solvated by water, whereas the N global minimum is only partially solvated. The Z-N enthalpy difference fluctuates as one proceeds to 16, 24 and 32 waters. The Boltzmann averaged MP2 $\Delta H_{N \to Z}$ predicted for 32 waters is 12.6, 13.3, or 11.5 kcal/mol using MP2(EFP1/DFT)//EFP2, MP2(EFP1/DFT) or PCM+MP2//MP2(EFP1/DFT), respectively. However, it is not clear that convergence has been attained at 32 water molecules. Therefore, systematic solvation from 33-49 waters is undertaken in this study.

An important consideration when exploring solvent effects with discrete solvent molecules is the large number of degrees of freedom that must be sampled in order to locate the global and all important low-lying minima. Monte Carlo (MC)[12] simulations are often employed to sample the configuration space. Such simulations can be very time-consuming since one often needs to explore millions of structures. In this work, the generalized effective fragment potential (EFP2)[13-17] method is employed to represent the alanine molecule in various conformations and the water molecules. Once the configura-

tion space has been sampled, a higher level of theory can be employed for the most important configurations. An efficient approach is to use quantum mechanics (QM) for the solute and a model potential (e.g., the EFP method) for the solvent. For these QM/MM calculations, water is described by the water-specific EFP1 method.[11,18,19]

The effective fragment potential (EFP1) and generalized effective fragment potential (EFP2)

EFP1 was designed specifically for water and is represented by a set of one-electron potentials that are added to the *ab initio* electronic Hamiltonian. The general approach is to derive two terms that represent Coulomb and induction (polarization) interactions and to then fit the remainder term to some *ab initio* level of theory. The Coulomb term is obtained from a Stone[20] distributed multipolar analysis, with the atom centers and the bond midpoints used as expansion points. The induction term is obtained from a set of distributed localized molecular orbital (LMO) polarizabilities. Therefore, the number of bonds and lone pairs equal the number of polarizability points. Many body effects are captured by iterating this dipole–induced dipole term to self-consistency. The multipoles and the polarizabilities are obtained from the chosen *ab initio* level of theory. Both of these terms are multiplied by damping (screening) expressions to account for near-field behavior.

One version of EFP1 is based on Hartree-Fock theory and is called EFP1/HF.[18] For EFP1/HF, the remainder term includes exchange repulsion and charge transfer. In the B3LYP[10] density functional theory version of EFP1 (EFP1/DFT),[11] the remainder term includes, in addition, short range correlation effects. The third version of EFP1 is based on second order Møller-Plesset perturbation theory (EFP1/MP2). EFP1/MP2[19] adds a

term that represents dispersion interactions, as well as second order corrections to the Coulomb and polarization terms. EFP1/HF and EFP1/DFT have both been interfaced with several levels of ground and excited state electronic structure theory, while EFP1/MP2 is currently only available for solvent-solvent, not solvent-solute, interactions and is not used in this study.

Recently, a more general version of EFP (EFP2) that is applicable to *any* molecular species has been developed.[13] EFP2 may be expressed in terms of five interaction energy terms:

$$E = E_{coul} + E_{ind} + E_{exrep} + E_{disp} + E_{ct}, \qquad (1)$$

The first two terms in Eq. (1), the Coulomb and induction interactions, are the same as those in EFP1. The remaining three terms represent interactions due to exchange-repulsion(E_{exrep}), dispersion(E_{disp}) and charge transfer(E_{ct}). Each of the terms in Eq. (1) is obtained from first principles with no empirically fitted parameters. The exchange repulsion is derived as an expansion in the intermolecular overlap. When this overlap expansion is expressed in terms of frozen LMOs on each fragment, the expansion can reliably be truncated at the quadratic term. This term does require each EFP to carry a basis set. Since the same basis set is used to generate the multipoles and the molecular polarizability tensor, EFP calculations are basis set dependent.

Dispersion interactions are often expressed by an inverse *R* expansion,

$$E_{disp} = \sum_n C_n R^{-n}, \qquad (2)$$

where the coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities integrated over the entire frequency range.[15] The first term in the expansion,

n=6, corresponds to the induced dipole-induced dipole (van der Waals) interactions. Damping of the dispersion term is formulated in terms of the LMO overlap.[16]

The charge transfer interaction is derived using a supermolecule approach, in which the occupied valence molecular orbitals on one fragment are allowed to interact with the virtual orbitals on another fragment.[17] This interaction term leads to significant energy lowering in *ab initio* calculations on ionic or highly polar species when incomplete basis sets are employed.

Computational methods

Gas phase results for N alanine were obtained with restricted Hartree-Fock[7] (RHF) and second order Møller-Plesset perturbation theory[6] (MP2) using the 6-31++G(d,p) basis set.[8] In order for the Z structures to be local minima in the absence of discrete water molecules, their structures must be optimized in the presence of a continuum solvent. The PCM approach[9] is used here for this purpose. The PCM uses van der Waals radii for cavity generation (see www.msg.chem.iastate.edu for details). The values used in these calculations may be found in the GAMESS manual. The Z structures were determined using the RHF/6-31++G(d,p) method, including the PCM continuum (denoted RHF+PCM). The N structures were obtained using the same combination of methods. To obtain more accurate energies, MP2+PCM single point calculations were performed at the RHF+PCM structures.

As discussed in the previous work,[5] the focus here is on the L-alpha alanine. Optimized "gas-phase" structures are illustrated in Figures 1 and 2 for the N and Z species, respectively. In Figure 2, several internal rotational Z structures are included. Since the EFP model has internally frozen geometries, this is an effective way to sample different Z

arrangements. This is useful for future reference when water molecules are added.

The Monte Carlo(MC)[12] with simulated annealing (SA)[21] method was used to sample the potential energy surface. It has been demonstrated that the EFP2 method provides reliable structures for amino acid – water clusters. Therefore, for computational efficiency, the MC simulations were carried out using the EFP2 method for both the solute and solvent. The molecular structures used to create the EFP2 for N and Z alanine were taken from RHF+PCM calculations. All eight forms of L-alanine shown in Figure 1 were used in the MC simulations. For Z alanine an EFP2 was generated for the L enatiomer and the three Z rotamers shown in Figure 2. The MC/SA method with local minimization was used to sample the configuration space. For each global minimum found, the number of structures sampled was on the order of 500,000 – 1,300,000. The number of steps taken for each temperature was varied from 100 to 10,000. The number of steps between local optimizations was varied from 10 to 1,000. The number of fragments moved per step was varied between one and five. The starting temperature for the simulated annealing varied from 500 to 20,000 K and the final temperature was 300 K. For each alanine isomer in the presence of a given number of waters, a minimum energy structure was found from MC/SA simulations. Additional higher energy structures were selected in 0.6 kcal/mol increments. Further optimizations and Hessians (matrices of energy second derivatives) were calculated at the EFP2 level of theory. From this set, a minimum energy structure was found and additional higher energy structures selected in steps of 1.2 kcal/mol. Then, after visual examination of the second subset, unique structures were optimized and verified to be potential energy minima by diagonalizing the corresponding Hessians, using MP2 for the alanine and EFP1/DFT for the water, MP2(EFP1/DFT), to

assess the accuracy of EFP2. Single points including PCM+MP2 were run at the MP2(EFP1/DFT) optimized geometries. All energies reported in the following sections have been corrected for the vibrational zero point energy. All calculations were performed with the general atomic and molecular electronic structure system (GAMESS),[22] which is freely available from Iowa State University. Structures were visualized with MacMolPlot,[23] a graphical interface to GAMESS.

Results

Alanine monomer

All neutral L alanine structures in Figure 1 were used in the Monte Carlo (MC) simulations. For the Z conformer of L alanine, the EFP2 conformational rotamers (45˚, 90˚ rotation of the COO^- group about the C-C bond, 120˚ rotation of NH_3^+ group about the N-C bond) in Figure 2 were used in the MC simulations. Previous work showed that the D chiral analogs did not significantly alter the relative N ➜ Z energy.[5]

Continuum Calculations

The HF+PCM/6-31++G(d,p) continuum method predicts the Z to be lower in energy than the N species; however, this difference is only 0.8 kcal/mol. The importance of including electron correlation has been demonstrated in previous studies.[3-5] A MP2+PCM/6-31++G(d,p) single point calculation at the HF+PCM geometry predicts the Z to be 1.8 kcal/mol lower in energy. MP2 may alter the gradient, and consequently, the geometry as well. When discrete waters are added to the system, the strengths of the hydrogen bonds formed between the solvent molecules and the N and Z isomers differ,[5] and the continuum solvent may not fully capture this differential effect. However, A combined discrete+continuum approach might result in a faster convergence of the Z-N en-

ergy difference.

In the following subsections the impact of increasing the number of water molecules on the N-Z equilibrium is examined, as are the detailed structures. The key Z and N structures and their relative energies are shown in Figures 3-9. A subset of these figures is discussed here. The remaining figures are available in Supplementary Material. The ordering of the structures in these figures is determined by the relative energies determined by the MP2+PCM/6-31++G(d,p) calculations, since MP2+PCM is considered to be the highest level of theory in this work.

Alanine(H$_2$O)$_{32}$ Figure 3

This previous paper on the aqueous solvation of alanine examined the addition of up to 32 water molecules[5]. This is the starting point for the further systematic addition of waters that is explored in this work. The lowest energy 32-water N isomer, N32-1, does not appear to be fully solvated. However, with 32 waters in the system, it is possible to find structures, such as N32-4, that appear to be fully solvated; that is completely surrounded by water molecules. At the highest level of theory, N32-4 is ~15 kcal/mol above N32-1, so the fully solvated species is clearly not the lowest energy N isomer. This may be due to the resistance of the nonpolar hydrocarbon moiety to aqueous solvation. The two lowest energy Z isomers, Z32-1 and Z32-2 are separated by only 2 kcal/mol, and both are more fully solvated than the N species. The two lowest energy N isomers are 9.1 and 9.6 kcal/mol, respectively, higher in energy than Z32-1 at the highest level of theory.

Alanine(H$_2$O)$_{33}$ Figure 4

The global minimum on the Z potential energy surface (Z33-1) results from an N isomer undergoing intramolecular hydrogen transfer when optimized with the alanine as

MP2 and the water as EFP1, MP2(EFP1/DFT)/6-31++G(d,p). This emphasizes the importance of *ab initio* optimization on a subset of species sampled from the MC simulations. The MC sampling of the zwitterion-$(H_2O)_{33}$ potential energy surface did find similar, energetically competitive (~1-3 kcal/mol higher in energy), structures compared to Z33-1. These structures also exhibit partial solvation, excluding the methyl group. N33-1 is a partially solvated structure 2.0 kcal/mol higher in energy than Z33-1 at the MP2(EFP1/DFT) optimized level of theory, and 9.7 kcal/mol higher when adding PCM in a single point calculation, MP2+PCM//MP2(EFP1/DFT)/6-31++G(d,p). Partially solvated clusters predominate the low-energy N isomers. For example, N33-2 and N33-3 are 8-12 and 10-20 kcal/mol above Z33-1 with MP2 and MP2+PCM, respectively. However there are collections of structures, represented by N33-4, in the energetic range of 20-30 kcal/mol above Z33-1 that may be considered fully solvated neutral isomers.

Alanine$(H_2O)_{36}$ Figure 5

In the 36-water lowest energy N minimum, N36-1, the methyl group is unsolvated, and the hydrogen bonding moieties, COOH and NH_2 are on the surface of a water cluster. In N36-2, about 1 kcal/mol higher in energy than N36-1, a water cluster forms around the methyl group with the COOH and NH_2 groups forming hydrogen bonds with the water molecules in the cluster, primarily as surface interactions. The Z36-1 – N36-1 energy difference, is ~13.6 kcal/mol. The PCM contributes ~1.5 kcal/mol to this energy difference. The addition of PCM has a large (9.4 kcal/mol) stabilizing effect on the energy of Z36-2 relative to Z36-1 stemming from a difference in the contribution of the hydrogen bonding pairs of COO^- and NH_3^+ surface areas to the PCM cavity. The higher energy N species N36-3 and N36-4 are both examples of mostly solvated neutral alanine

isomers that are 15-20 kcal/mol higher in energy than the lowest energy N structure.

Alanine(H₂O)₄₁ Figure 6

In Z41-1 and Z41-2, one carboxylate oxygen atom participates in a surface hydrogen bond. This surface hydrogen bond appears to be correlated with large (>3 kcal/mol) changes in relative energies when using PCM+MP2 compared to MP2. N41-1 is 15.7 (20.0) kcal/mol higher in energy relative to Z41-1 at the MP2 (MP2+PCM) level of theory. In N41-1 the water forms a large ring around the methyl, mostly solvating the alanine. Such structures typically lie ~20 kcal/mol higher in energy that the global minimum. N41-2 is only slightly higher in energy than N41-1 and has the alanine adjacent to and forming hydrogen bonds with the water cluster, with hydrogen via the COOH and NH₂ groups. N41-3 and N41-4 are both fully solvated, and 4 and 7 kcal/mol, respectively above N41-1. N41-3 exhibits a solvation sphere around the methyl, whereas the N41-4 has an elongated water cluster that is roughly perpendicular to the COOH and NH₂ moieties.

Alanine(H₂O)₄₆ Figure 7

For 46 water molecules, the two lowest energy Z structures, Z46-1 and Z46-2, appear to be fully solvated. N46-1 is the first example of a fully solvated N species that is also the lowest energy N isomer, ~12 kcal/mol above Z46-1 at the highest level of theory. The Z-N enthalpy differences are within the expected range, based on the glycine $\Delta H_{N \to Z}$ of 10.3 kcal/mol.[2] The Z isomers are separated by 1.3 kcal/mol at the same level of theory. N46-1 and N46-3 appear at first glance to be similar; however, N46-1 has more hydrogen bonds (defined as an O-O distance less than 3 Å and an OOH angle less than 150°). In N46-2 the alanine forms surface hydrogen bonds with a water cluster; similar

structures are found as the N lowest energy minima for smaller numbers of water molecules.

Alanine(H_2O)$_{48}$ Figure 8

The two lowest energy Z structures, Z48-1 and Z48-2, differ by only 2.4 kcal/mol at the highest level of theory and have similar, fully solvated structures. In Z48-1 the water molecules that sit below the NH_3^+ group are arranged approximately in a prism structure, while in Z48-2 these six waters are arranged in a chair-like conformation. In both cases, Z48-1 and Z48-2, these six waters appear to the beginning of a second solvent shell. Z48-1 is 10.5 kcal/mol lower in energy than N48-1, the N global minimum, at the highest level of theory. All of the N isomers presented in Figure 19 are fully solvated. With 48 water molecules present, partially solvated structures have become less common in the MC simulations. The N alanine isomers 1L and 2L (Figure 1) predominate in the lower energy structures.

Alanine(H_2O)$_{49}$ Figure 9

The two lowest 49-water Z species, Z49-1 and Z49-2, differ in energy by 3.3 kcal/mol at the highest level of theory, MP2+PCM. The structures show an expanded region of water molecules around the methyl group. It appears that the water is now truly beginning to form a second solvation sphere. Z49-1 is 11.5 kcal/mol lower in energy than N49-1. Most low-energy N structures are fully solvated, although there are still populations of separated alanine and water clusters exhibited by N49-2, which are ~15-20 kcal/mol higher in energy than Z49-1. N49-3 and N49-4 represent higher energy solvated isomers.

Summary

Table 1 presents a summary of the Z-N 0˚K enthalpy differences discussed in the previous sub-sections, as well as the temperature corrected enthalpies at 298.15˚K. The zero point vibrational corrections and the temperature corrections were obtained using the harmonic oscillator and rigid rotor approximations. All of the enthalpy differences presented in Table 1 oscillate as the number of water molecules is increased, especially for n = 32 until n = 46. From n = 46 to n = 49, these oscillations are much smaller, ranging from ~10 kcal/mol to ~12.5 kcal/mol. This is in the range of the anticipated enthalpy difference for alanine. The temperature corrections to 298K have little impact on the computed values. The point at which the two alanine species, Z and N, appear to become fully solvated is not the same. Starting from 42 water molecules, the global Z minimum appears to be fully solvated, while this point only occurs at 46 water molecules for the N global minimum.

Boltzmann averaging

Boltzmann averaged energies were calculated for each water cluster using the Boltzmann equation:

$$\frac{\sum_i X_i e^{-\Delta E_i / RT}}{\sum_i e^{-\Delta E_i / RT}} = X_n \ , \tag{3}$$

where X_i is the calculated property (e.g., the energy) of the *i*th structure, corrected with zero point vibrational energy. ΔE_i is calculated by taking the difference between the energy of the ith structure and the energy of the lowest energy structure for a given n; T = 273 K. X_n is the Boltzmann averaged value of X for all unique structures determined by MP2(EFP1/DFT)/6-31++G(d,p) optimizations for a given number (n) of waters. Boltzmann averaged Z-N energies are presented in Table 2. These enthalpies do not change

from those reported in Table 1 by more than 1 kcal/mol.

Correlation energy and continuum solvation

Table 3 presents the effects of correlation energy and continuum solvation on the Z-N 0˚K enthalpy differences. The correlation energy contribution is determined by subtracting the MP2(EFP1/DFT)/6-31++G(d,p) enthalpy from the RHF(EFP1/DFT)/6-31++G(d,p) enthalpy. The continuum energy contribution is determined by subtracting the MP2+PCM//MP2(EFP1/DFT)/6-31++G(d,p) enthalpy from the MP2(EFP1/DFT)/6-31++G(d,p) enthalpy. The Boltzmann averaged estimates of the correlation and continuum corrections are very similar to those obtained from the simple differences between the respective global minima. The two corrections oscillate somewhat as the number of water molecules present increases. The oscillations in the correlation corrections are smaller than those due to the continuum corrections. Once n = 46, the latter oscillations decrease as well. In this range the both corrections are ~2 kcal/mol or less. The N global minimum is fully solvated by the addition of 46 waters, and alanine does not significantly contribute the electrostatic stabilization of the PCM cavity. This may be part of the reason why the PCM oscillations with 46 or more water molecules greater are smaller. The approach to complete solvation appears to be neither monotonic nor linear. In addition, as the number of water molecules is increased, the complexity of determining the global minimum increases, the exploration of millions of configurations notwithstanding. Both of these factors very likely contribute to the oscillations seen in Tables 1-3.

Conclusions

The solvation of alanine by water, and the alanine enthalpy change for the N ➜ Z isomerization ($\Delta H_{N \to Z}$), has been studied by systematically adding 32-49 waters. At ap-

proximately 42 waters, the Z alanine - water complex becomes consistently fully solvated. Starting at 45 waters, the solvent structure varies predominately in the placement of a few surface water molecules. Therefore, 41-45 waters may be thought of as the first solvent shell. The additional waters beyond this are relatively free to move on the outside of the shell, beginning a second solvation shell. Fully solvated N alanine arrangements appear at 32 waters, but these structures are significantly higher in energy than the N minimum. At about 42 water molecules, fully solvated N alanine structures begin to be energetically competitive with the partially solvated global N minimum. At 46 waters both the Z and N global minima are fully solvated. This also begins the stabilization of the $\Delta H_{N \rightarrow Z}$.

Initial studies[5] of alanine solvated by PCM alone yield a $\Delta H_{N \rightarrow Z}$ of 1.8 kcal/mol, which is in qualitative, but not quantitative agreement with the anticipated experimental value. PCM does not significantly change the relative Z-N energies for water structures which completely solvate the alanine molecule. Upon the addition of 46 waters both the N global minimum and Z global minimum are fully solvated. The contributions from PCM are then small, ~ 2 kcal/mol. The correlation contributions to the 0K enthalpies are smaller than the continuum solvent contributions. Correlation energy may effect the identification of global minima for a given number of water molecules (e.g., n = 44). Therefore, correlation energy corrections must be included for a complete description of the $\Delta H_{N \rightarrow Z}$.

The trends and the absolute Z-N enthalpy differences are essentially the same for the 0K and 298K differences in the global minima, as well as for the 0K Boltzmann averages. Significant oscillation in these differences is observed until about 46 water mole-

cules, whereupon the enthalpy difference settles down to ~10-12 kcal/mol. The oscillations in these enthalpy differences are ascribed to the lack of monotonic convergence to fully solvated species for the Z and N species and to the difficulty in determining global minima with large numbers of water molecules. Nonetheless, the enthalpy difference appears to converge to a value that is consistent with experimental expectations. The $\Delta H_{N \to Z}$ of glycine is 10.3 kcal/mol,[2] and one would expect alanine to have a similar value. It appears that the alanine value is converged to within ~2.5 kcal/mol. There is no reason to assume $\Delta H_{N \to Z}$ converges quickly.

Acknowledgments

References:

1. (a)Albrecht, G.; Corey, R. B. *J. Am. Chem. Soc.* **1939**, *61*, 1087. (b) Jonsson, P. G.; Kvick, A. *Acta Crystallogr. Sect. B* **1972**, *28*, 1827.
2. Wada, G.; Tamura,E.; Okina, M.; Nakamura, M. *Bul. Chem. Soc. Jpn.* **1982** , *55*, 3064.
3. Aikens, C. M.; Gordon, M. S. *J. Am. Chem. Soc.* **2006**, *128*, 12835.
4. Chuchev, K.; BelBruno,J. J. *J.Mol. Struct.: THEOCHEM.* **2008**, *850*, 111.
5. Mullin, J. M.; Gordon, M. S. *J. Phys Chem. Submitted.*
6. Moller, C.; Plesset, S. *Phys. Rev.* **1934**, *46*, 618.
7. Roothan, C. C. *J. Rev. Mod. Phys.* **1960**, *32*, 179.
8. (a) Hehre, W.J.; Ditchfield, R.; Pople, J.A. *J. Chem. Phys.* **1972**, *56*, 2257. (b) Francl, M.M.; Pietro, W.J.; Hehre, W.J.; Binkley, J.S.; Gordon, M.S.; DeFrees, D.J.; Pople, J.A. *J.Chem.Phys.* **1982**, *77*, 3654. (c) Hariharan, P.C.; Pople, J.A. *Theoret.Chim.Acta* **1973**, *28*, 213.
9. (a) Miertus, S.; Scrocco, E.; Tomasi, J. *J. chem. phys.* **1981**, *55*, 117. (b) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 30323041. (c) Li, H.; Jensen, J.H. *J. Compu. Chem.* **2004**, *25*, 144.
10. (a) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098. (b) Lee, C. C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
11. Adamovic, I.; Freitag, M. A.; Gordon, M. S. *J. Chem. Phys.* **2003**, *118*, 6725.

12. (a) Metropolis, N.; Rosenbluth, A.; Teller, A. *J. Chem. Phys.* **1953**, *21*, 1087. (b) Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill, G.N. *J.Chem.Phys.* **2000**, *112*, 2063.

13. (a)Gordon, M.S.; Slipchenko, L.; Hui, L.; Jensen, J.H. *Ann. Rep. Comp. Chem.* **2007**, *3*, 177. (b) Gordon, M.S.; Freitag, M.A.; Bandyopadhyay, P.; Kairys, V.; Jensen, J.H.; Stevens, W.J. *J. Phys. Chem.* **2001**, *105*, 293.

14. (a) Slipchenko, L.; Gordon, M.S. *J. Comp. Chem.* **2006**, *28*, 276. (b) Freitag, M.A.; Gordon, M.S.; Jensen, J.H.; Stevens W.J. *J. Chem. Phys.* **2000**, *112*, 7300.

15. (a) Adamovic, I.; Gordon, M.S. *Mol. Phys.* **2005**, *103*, 379. (b) Amos, R.D.; Handy, N.C.; Knowles, P.J.; Rice, J.E.; Stone, A.J. *J. Phys. Chem.* **1985**, *89*, 2186. (c) Piecuch, P. *Mol. Phys.* **1986,** 59, 1085.

16. (a) Slipchenko, L.; Gordon, M. S. *J. Mol. Phys. Submitted.* (b) Murrell, J. N.; Teixeira-Dias, J. J. C. *Mol. Phys.* **1970**, *19*, 521.

17. Li. H.; Jensen, J.H. Gordon, M.S. *J. Chem. Phys.* **2006**, *124*, 214108.

18. (a) Jensen, J. H.; Gordon, M. S. *J. Chem. Phys.* **1996**, *104*, 7795. (b) Jensen, J. H.; Gordon, M. S. *Mol. Phys.* **1996**, *89*, 1313. (c) Jensen, J. H.; Gordon, M. S. *J. Chem. Phys.* **1998**, *108*, 4772. (d) Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, *105*, 11081. (e) Day, P.N.; Jensen, J.H.; Gordon, M.S.; Webb, S.P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.

19. Song, J.; Gordon, M. S. unpublished.

20. (a) Stone, J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047. (b) Stone, A. J. The Theory of Intermolecular Forces Oxford University Press, London, UK, 1996.

21. (a)Kirkpatrick, S.; Gelatt, C. D.; M. P. Vecchi *Science* **1983**, *220*, 671. (b) Cerny, V. *J. Optimization Theory and Applications* **1985**, *45*, 41.

22. (a) 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.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem.* **1993**, *14*, 1347. (b) Gordon, M. S.; Schmidt, M. W., Theory and Applications of Computational Chemistry, the first forty years; Elsevier; Amsterdam; 2005.

23. Bode, B. M.; Gordon, M. S. *J. Mol. Graphics Mod.* **1998**, *16*, 133.

Figure Captions

Figure 1. L neutral alanine conformer structures and energies (relative to ALA 1 L) at RHF/6-31++G(d,p), MP2/6-31++G(d,p), PCM+RHF/6-31++G(d,p), PCM+RHF/6-31++G(d,p) levels of theory.

Figure 2. Alanine zwitterion and the three rotamers (45°, 90° rotation of the COO^- group about the C-C bond, 60° rotation of NH_3^+ about the N-C bond) used in Monte Carlo.

Figure 3. Alanine and 32 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 4. Alanine and 33 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 5. Alanine and 36 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 6. Alanine and 41 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global

minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 7. Alanine and 46 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 8. Alanine and 48 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure 9. Alanine and 49 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Table 1.

0˚K and 298.15˚K Enthalpy of N-Z transition with MP2(EFP1/DFT)/6-31++G(d,p) opti-mization and MP2+PCM//MP2(EFP1/DFT)/6-31++G(d,p) levels of theory. Zero point energy corrected.

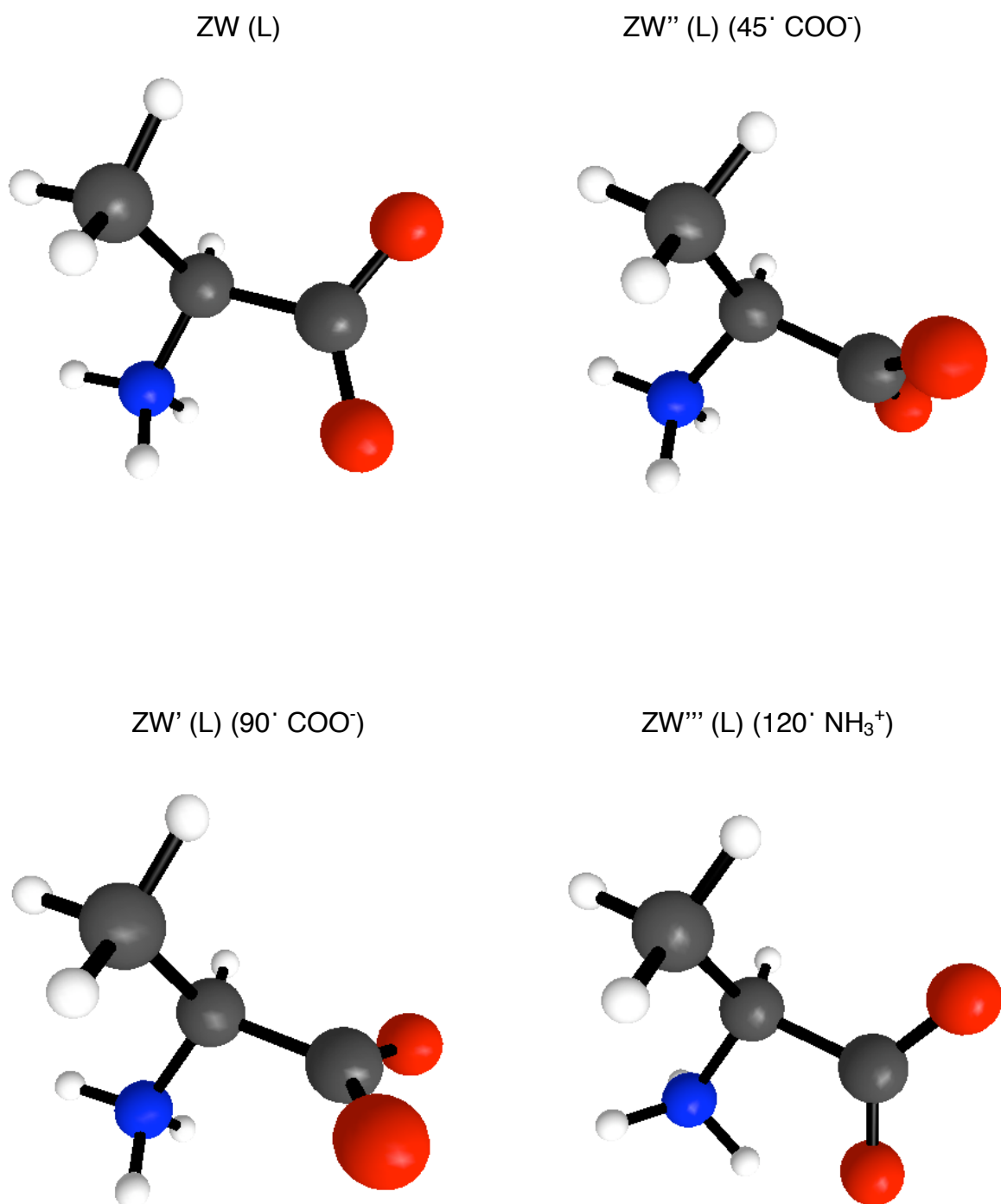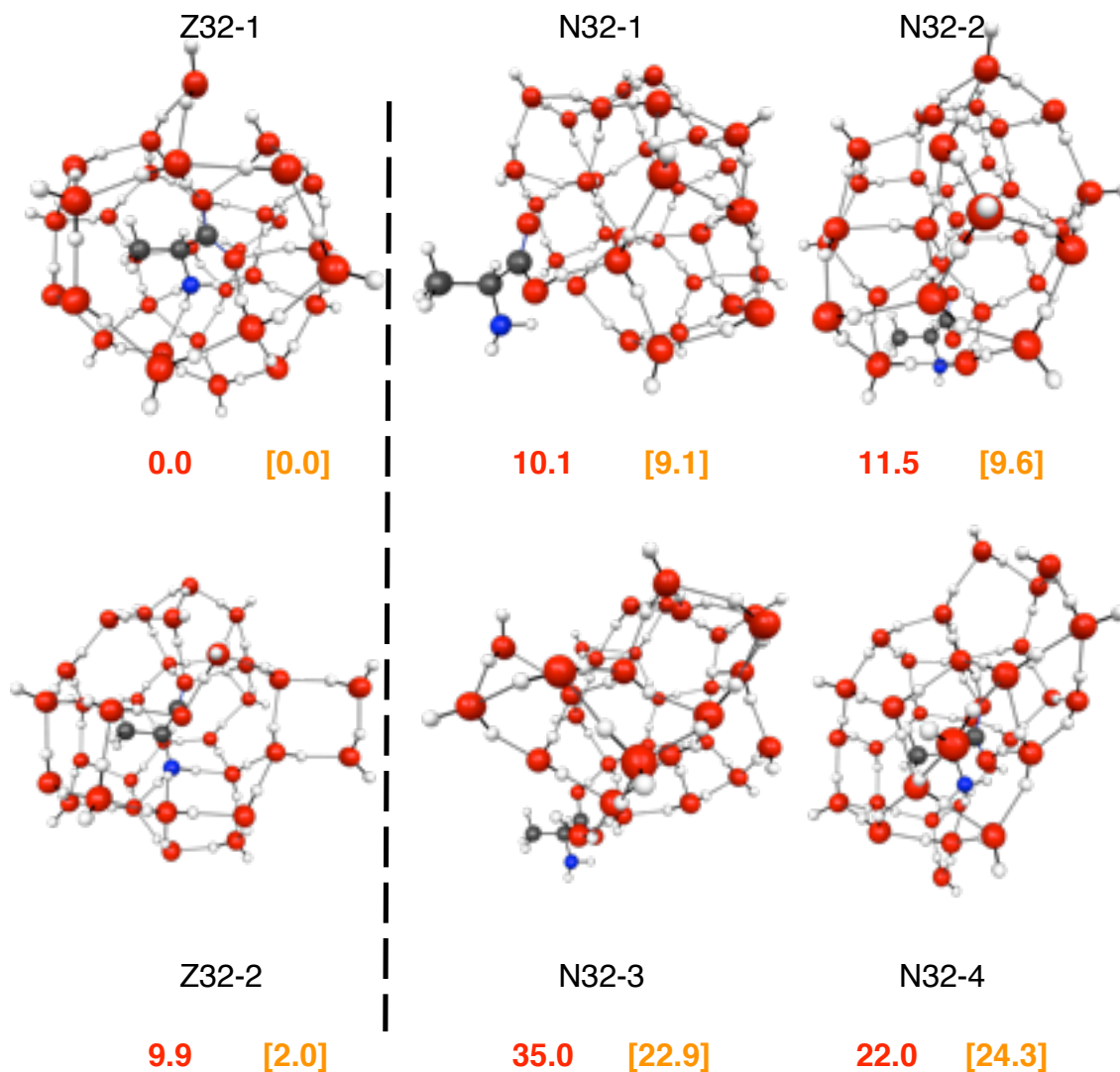|	0˚K Enthalpy		298.15˚K Enthalpy	
	MP2(EFP1/DFT)*	MP2+PCM//MP2(EFP1/DFT)*	MP2(EFP1/DFT)*	MP2+PCM//MP2(EFP1/DFT)*
$(H_2O)_n$	Z-N	Z-N	Z-N	Z-N
32	10.1	9.1	12.2	13.1
33	2.0	9.7	6.1	13.8
34	11.7	15.6	11.6	16.2
35	6.0	8.5	4.7	7.3
36	12.1	13.6	11.1	12.6
37	7.0	3.6	6.9	3.5
38	12.3	13.2	12.4	13.3
39	6.0	10.6	6.0	10.5
40	10.7	11.4	11.5	12.0
41	15.7	20.0	15.4	19.6
42	5.8	6.8	5.9	6.9
43	10.8	8.8	9.8	8.5
44	15.7	8.1	15.4	7.8
45	5.4	6.9	5.2	6.8
46	13.1	12.1	13.2	12.3
47	10.1	12.6	10.0	12.6
48	12.2	10.5	11.8	10.1
49	12.2	11.6	12.5	11.9

* 6-31++G(d,p) ZPE Corrected

Table 2.

Boltzmann averaged 0˚K and 298.15˚K Enthalpy of N-Z transition for MP2(EFP1/DFT)/6-31++G(d,p) optimization and MP2+PCM//MP2(EFP1/DFT)/6-31++G(d,p) levels of theory. Zero point energy corrected.

| | 0˚K Enthalpy | | 298.15˚K Enthalpy | |
| | MP2(EFP1/DFT)* | MP2+PCM// MP2(EFP1/DFT)* | MP2(EFP1/DFT)* | MP2+PCM// MP2(EFP1/DFT)* |
$(H_2O)_n$	Z-N	Z-N	Z-N	Z-N
32	10.1	9.2	12.3	13.3
33	2.0	9.8	6.1	13.8
34	11.7	15.7	11.6	16.4
35	6.0	8.5	4.7	7.3
36	12.3	13.8	11.1	12.6
37	7.1	3.6	7.1	3.5
38	12.4	13.2	12.4	13.3
39	6.0	10.7	6.0	10.6
40	10.7	11.6	11.5	12.1
41	15.8	20.1	15.4	19.8
42	5.8	6.8	5.9	6.9
43	10.8	8.8	9.8	8.5
44	15.7	8.1	15.5	7.8
45	5.2	6.8	5.2	6.8
46	13.1	12.0	13.2	12.3
47	10.1	12.6	10.0	12.6
48	12.2	10.5	11.8	10.1
49	12.2	11.6	12.5	11.9

* 6-31++G(d,p) ZPE Corrected

Table 3.

Effect of including electron correlation and the continuum solvent. Both non-Boltzmann averaged (i.e., Z-N global minimum 0K enthalpy differences) and Boltzmann averaged 0˚K enthalpy differences are presented. The correlation energy contribution is MP2(EFP1/DFT)/6-31++G(d,p) subtracted from RHF(EFP1/DFT)/6-31++G(d,p). The continuum energy contribution is MP2+PCM//MP2(EFP1/DFT)/6-31++G(d,p) subtracted from MP2(EFP1/DFT)/6-31++G(d,p).

	non-Boltzmann averaged	Boltzmann averaged	non-Boltzmann averaged	Boltzmann averaged
$(H_2O)_n$	Correlation		Continuum	
32	-0.6	-0.6	-1.0	-1.1
33	1.4	1.4	-7.7	-7.8
34	-0.1	-0.1	-3.9	-3.9
35	1.3	1.3	-2.5	-2.5
36	-1.3	-1.3	-1.5	-1.5
37	3.7	3.8	3.4	3.5
38	0.7	0.7	-0.8	-0.8
39	2.1	2.2	-4.6	-4.6
40	0.1	0.1	-0.7	-0.9
41	-2.5	-2.6	-4.2	-4.2
42	3.2	3.2	-1.0	-1.0
43	0.1	0.1	1.5	1.5
44	-2.4	-2.5	7.6	7.7
45	-1.0	-1.0	-1.5	-1.6
46	-1.7	-1.9	1.0	1.1
47	-1.3	-1.3	-2.6	-2.6
48	-0.7	-0.7	1.7	1.8
49	-0.6	-0.7	0.6	0.6

Figure 1. L form neutral alanine conformer structures and energies(relative to ALA 1 L) at |RHF/6-31++G(d,p)|, MP2/6-31++G(d,p), {RHF+PCM/6-31++G(d,p)}, [MP2+PCM/6-31++G(d,p)//MP2/6-31++G(d,p)] levels of theory.
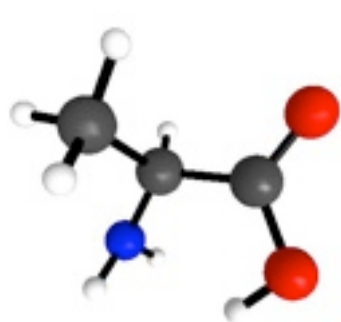
ALA 7L |5.4| 6.4   {3.8}   [4.5]   ALA 8L |5.8| 6.8   {4.1}   [4.9]
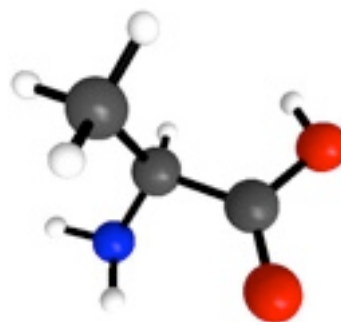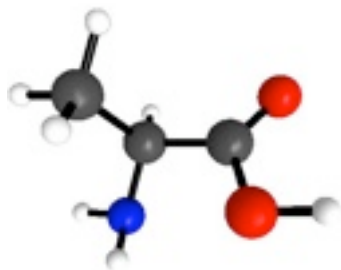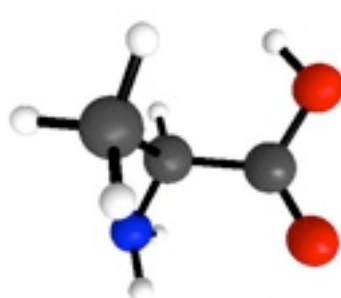
ALA 5L |-0.5| 0.3   {1.9}   [2.3]   ALA 6L |4.2| 5.6   {3.2}   [4.1]
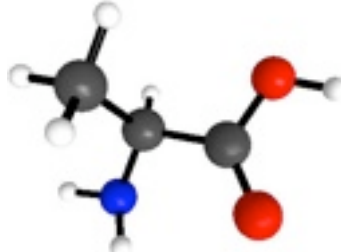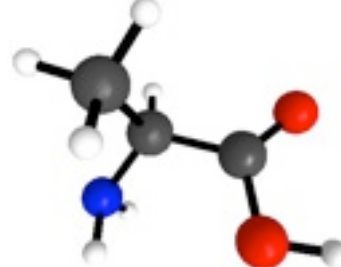
ALA 5L |-1.2| 0.6   {1.3}   [0.4]   ALA 6L |-1.5| 0.3   {1.3}   [2.4]
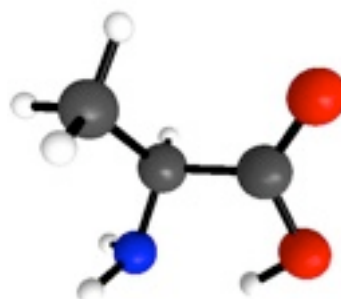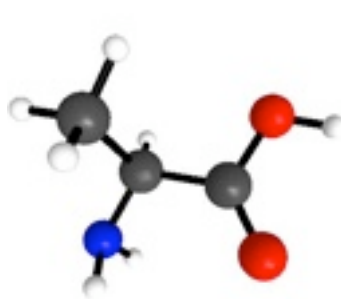
ALA 5L |0.0| 0.0   {0.0}   [0.0]   ALA 6L |-2.8| -0.7   {0.4}   [-0.7]

Figure 2. Alanine zwitterion and the three rotamers used in Monte Carlo.

ZW (L)

ZW" (L) (45˙ COO⁻)
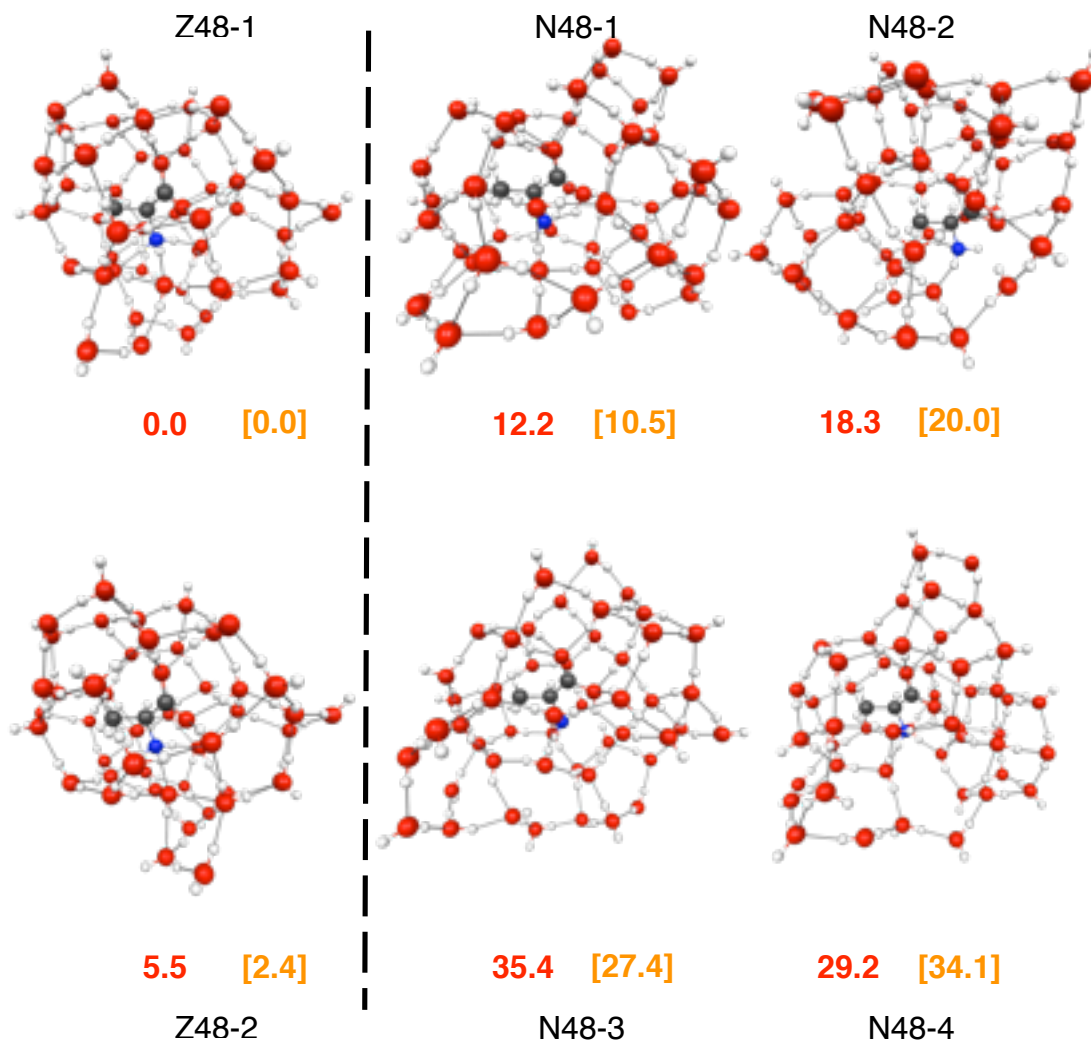

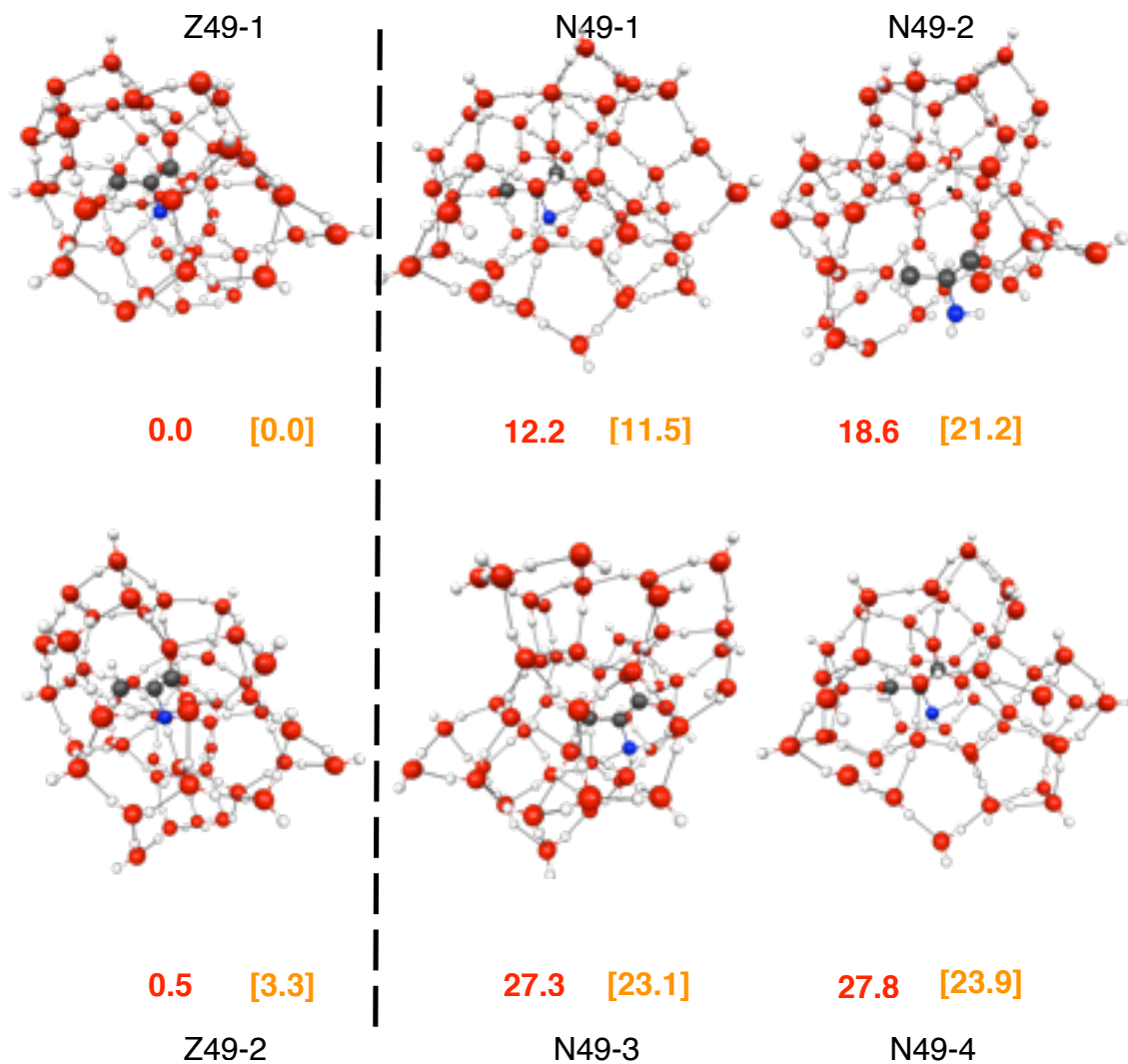ZW' (L) (90˙ COO⁻)

ZW''' (L) (120˙ NH₃⁺)

Figure 3. Alanine and 32 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z32-1	N32-1	N32-2
0.0 [0.0]	10.1 [9.1]	11.5 [9.6]

Z32-2	N32-3	N32-4
9.9 [2.0]	35.0 [22.9]	22.0 [24.3]

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 4. Alanine and 33 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z33-1   N33-1   N33-2

0.0   [0.0]   2.0   [9.7]   8.0   [11.3]

6.5   [9.2]   9.6   [19.8]   15.6   [23.8]

Z33-2   N33-3   N33-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 5. Alanine and 36 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z36-1

N36-1

N36-2

0.0   [0.0]

12.1   [13.6]

13.1   [14.5]

16.2   [6.8]

28.3   [27.5]

24.9   [31.6]

Z36-2

N36-3

N36-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 6. Alanine and 41 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 7. Alanine and 46 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z46-1

0.0   [0.0]

N46-1

13.1   [12.1]

N46-2

18.2   [21.1]

2.5   [1.3]

Z46-2

18.2   [24.4]

N46-3

21.3   [24.7]

N46-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 8. Alanine and 48 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z48-1

N48-1

N48-2

0.0   [0.0]

12.2   [10.5]

18.3   [20.0]

5.5   [2.4]

35.4   [27.4]

29.2   [34.1]

Z48-2

N48-3

N48-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure 9. Alanine and 49 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
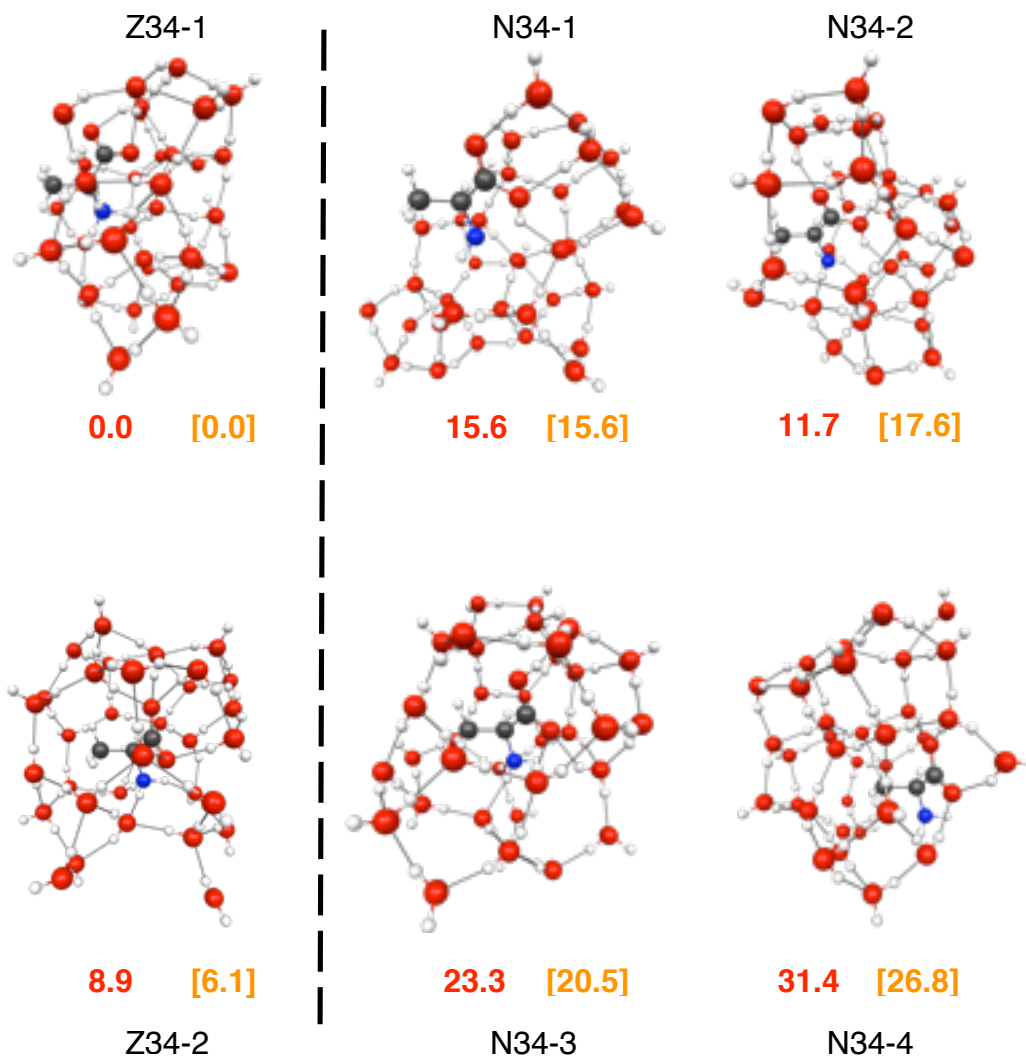

Z49-1

0.0   [0.0]

N49-1

12.2   [11.5]

N49-2

18.6   [21.2]

0.5   [3.3]

Z49-2

27.3   [23.1]

N49-3

27.8   [23.9]

N49-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

SUPPLEMENTARY MATERIAL
(CHAPTER 3. ALANINE: FROM PUDDLES TO PONDS**)**

Figure S1. Alanine and 34 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S2. Alanine and 35 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S3. Alanine and 37 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S4. Alanine and 38 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
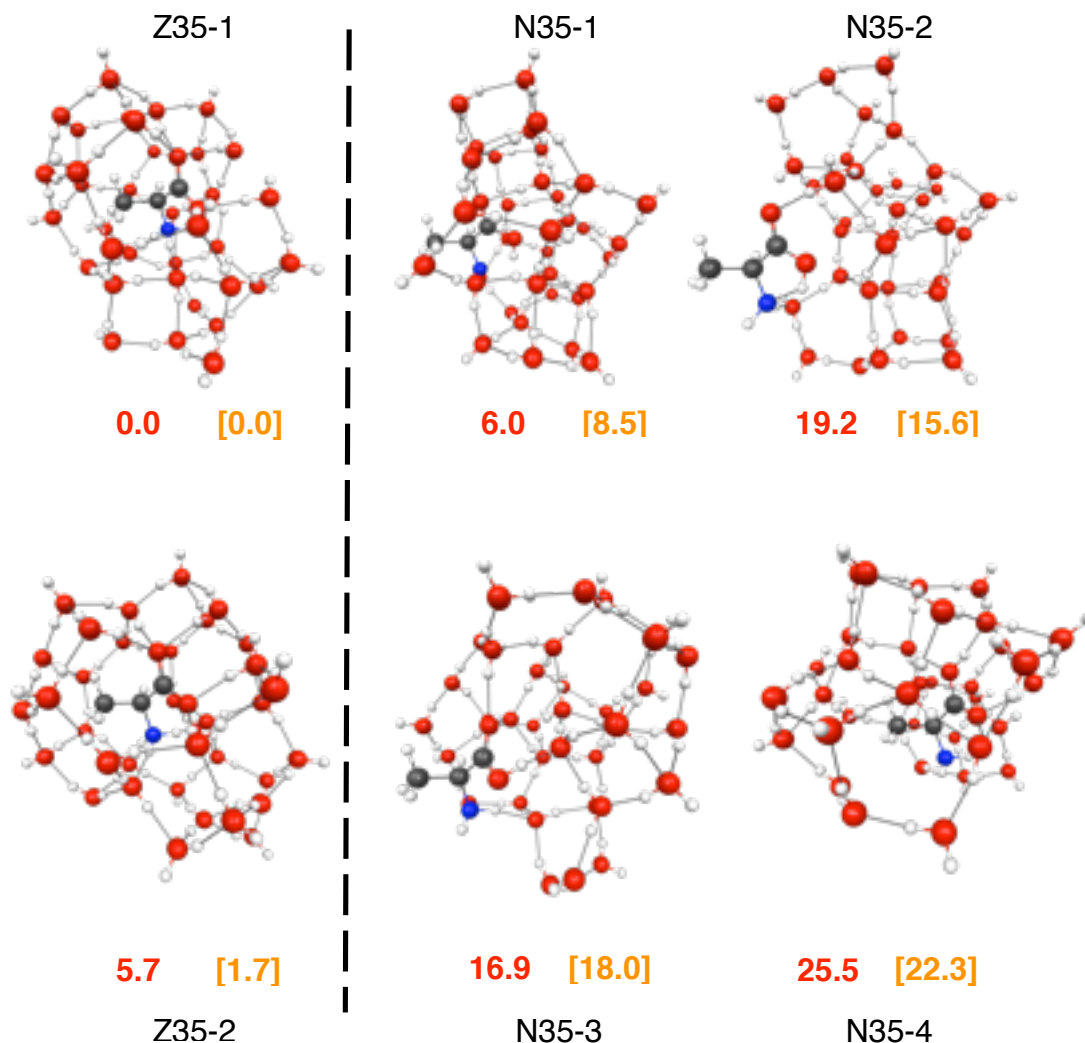
Figure S5. Alanine and 39 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with rep-

resentative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
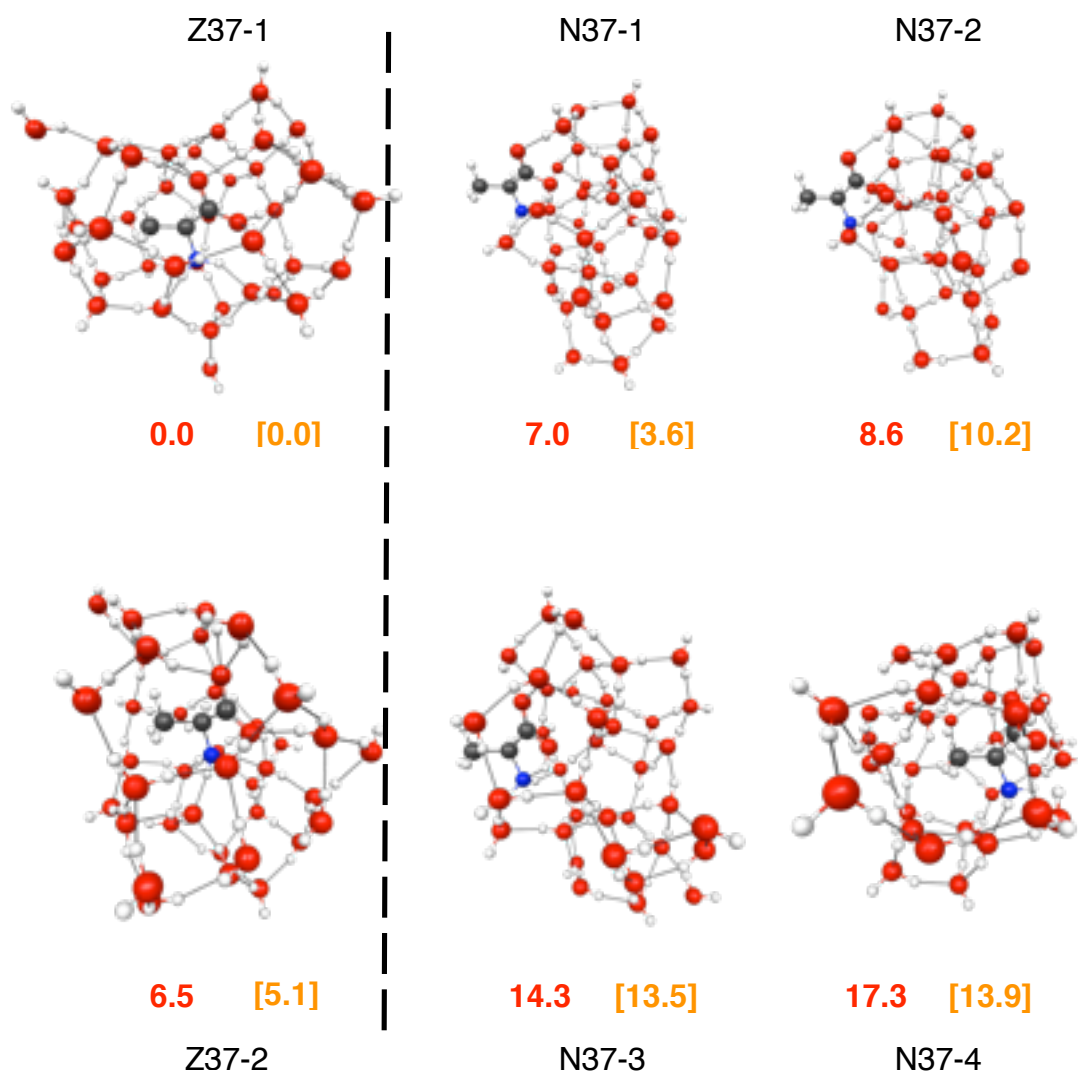
Figure S6. Alanine and 40 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S7. Alanine and 42 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
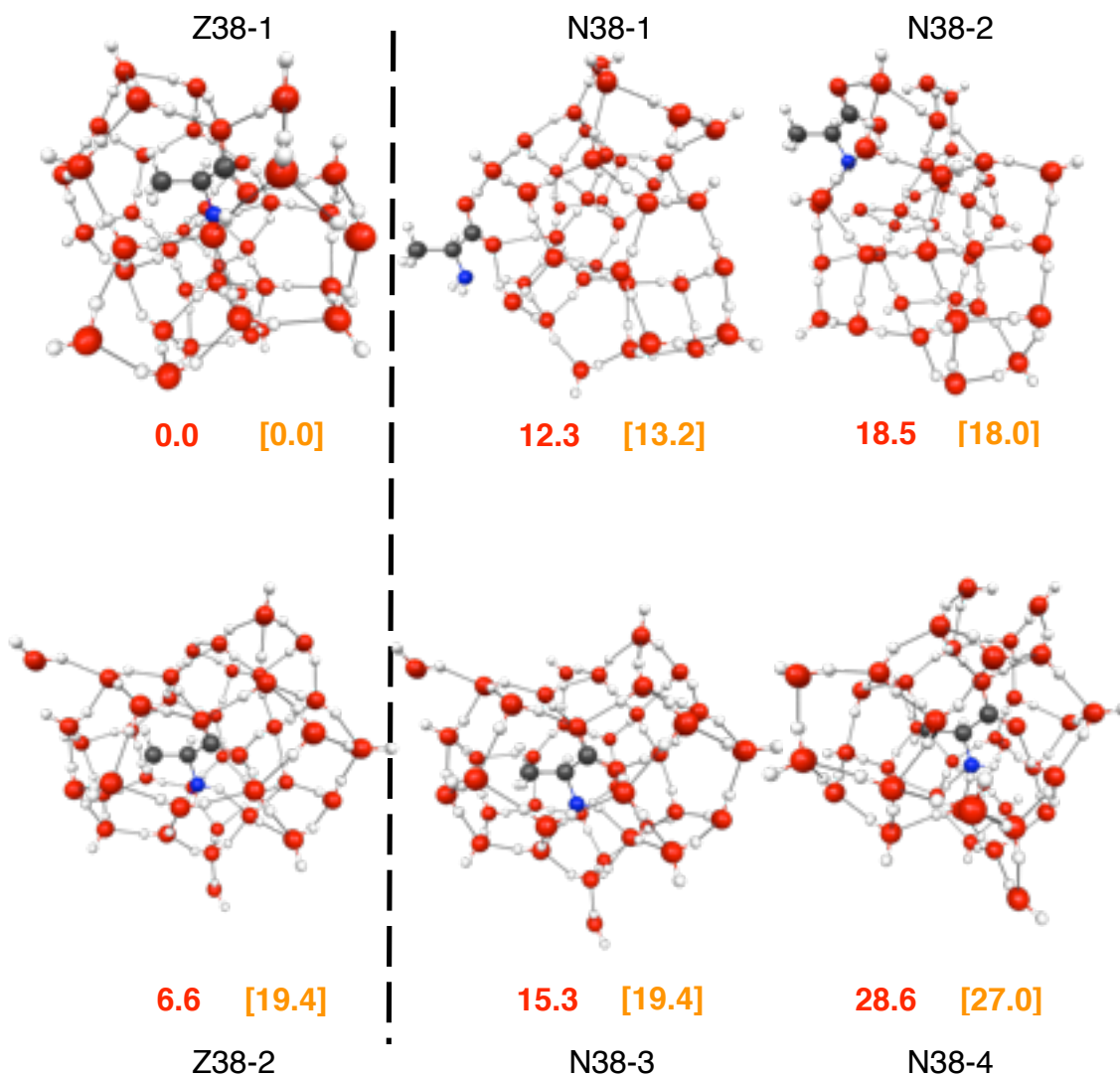
Figure S8. Alanine and 43 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S9. Alanine and 44 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
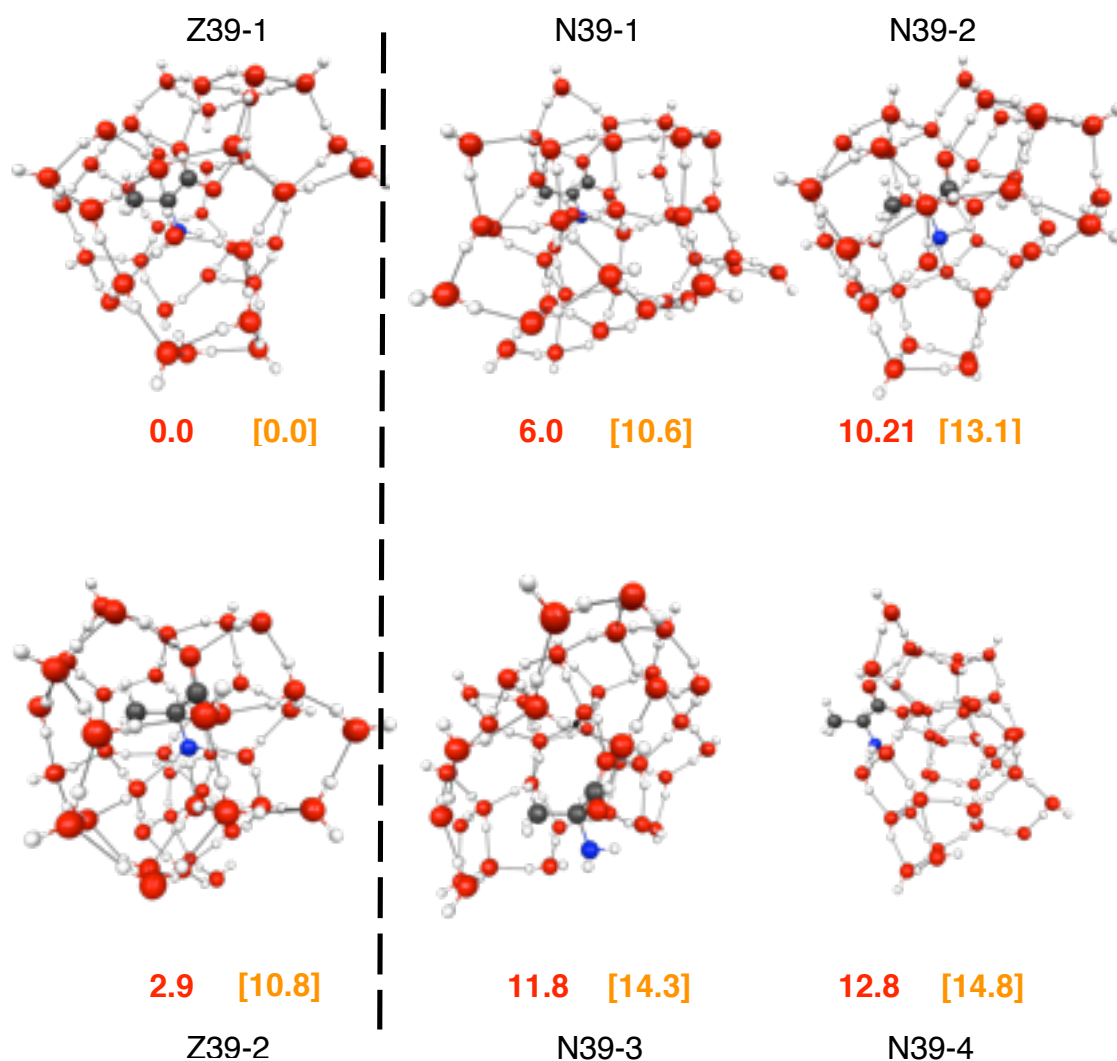
Figure S10. Alanine and 45 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).
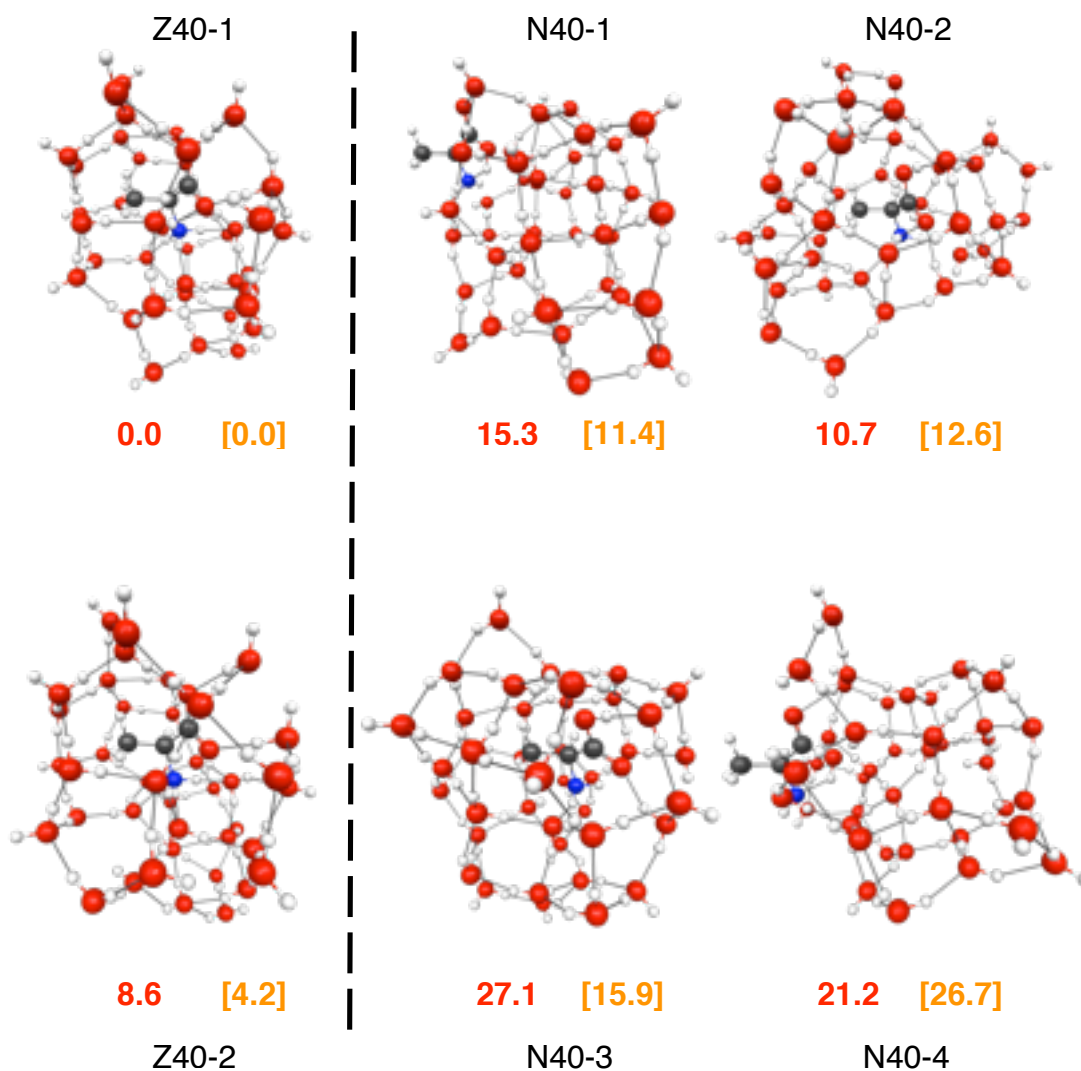
Figure S11. Alanine and 47 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Figure S1. Alanine and 34 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z34-1

N34-1

N34-2

0.0   [0.0]

15.6   [15.6]

11.7   [17.6]

8.9   [6.1]

23.3   [20.5]

31.4   [26.8]

Z34-2

N34-3

N34-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
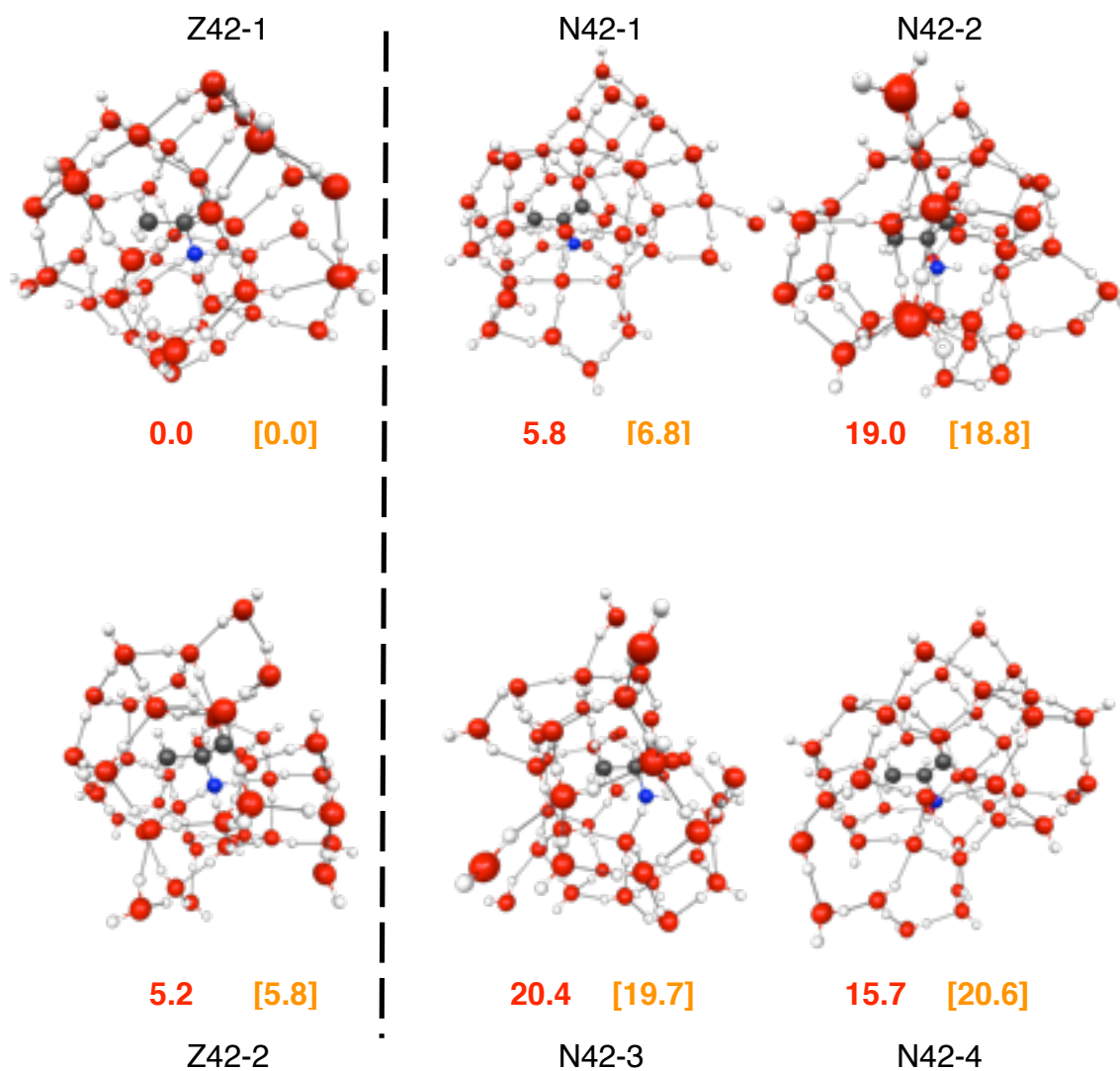PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S2. Alanine and 35 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z35-1

N35-1

N35-2

0.0   [0.0]

6.0   [8.5]

19.2   [15.6]

5.7   [1.7]

16.9   [18.0]

25.5   [22.3]

Z35-2

N35-3

N35-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
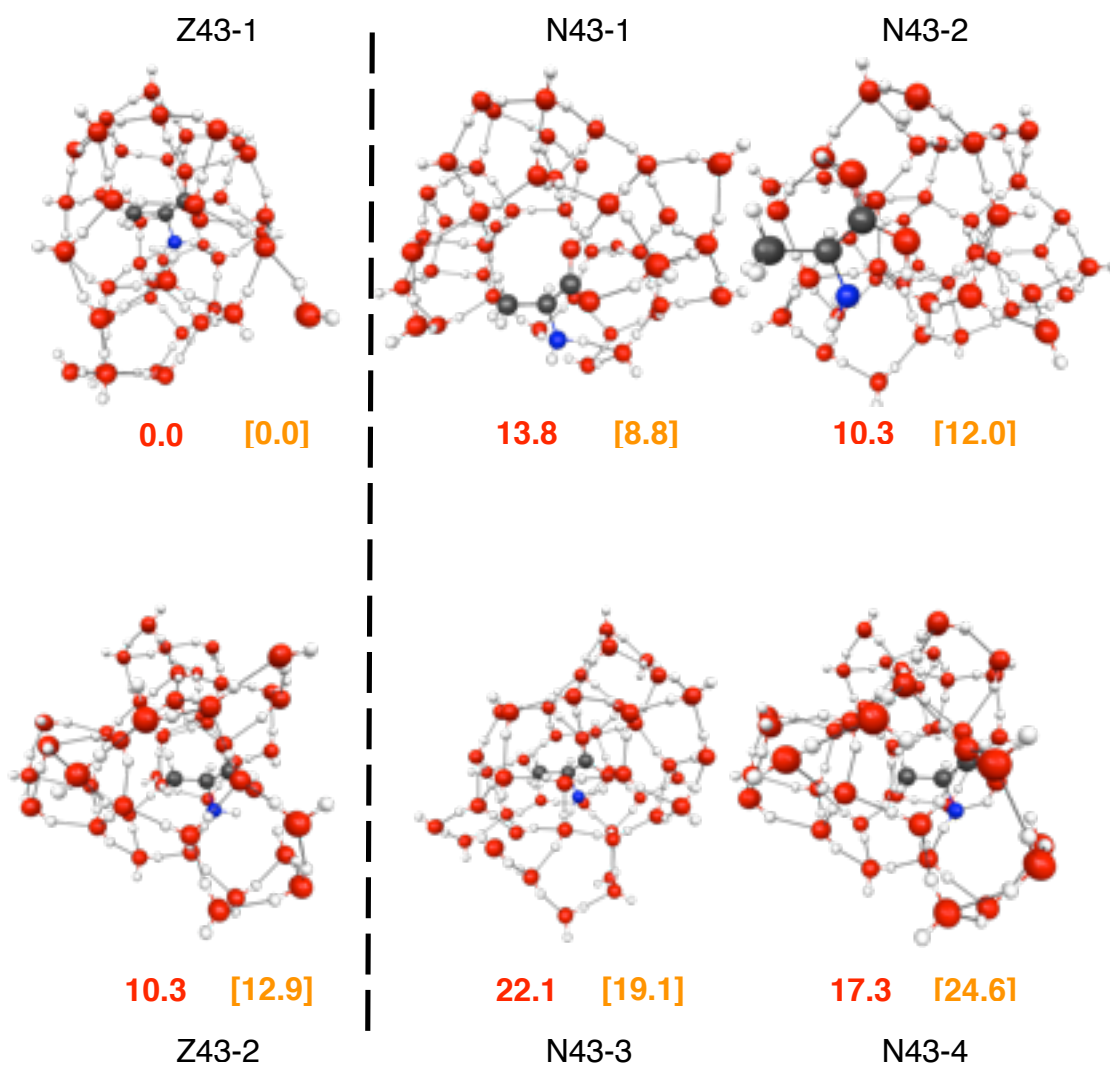PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S3. Alanine and 37 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z37-1   N37-1   N37-2

0.0   [0.0]   7.0   [3.6]   8.6   [10.2]

6.5   [5.1]   14.3   [13.5]   17.3   [13.9]

Z37-2   N37-3   N37-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)
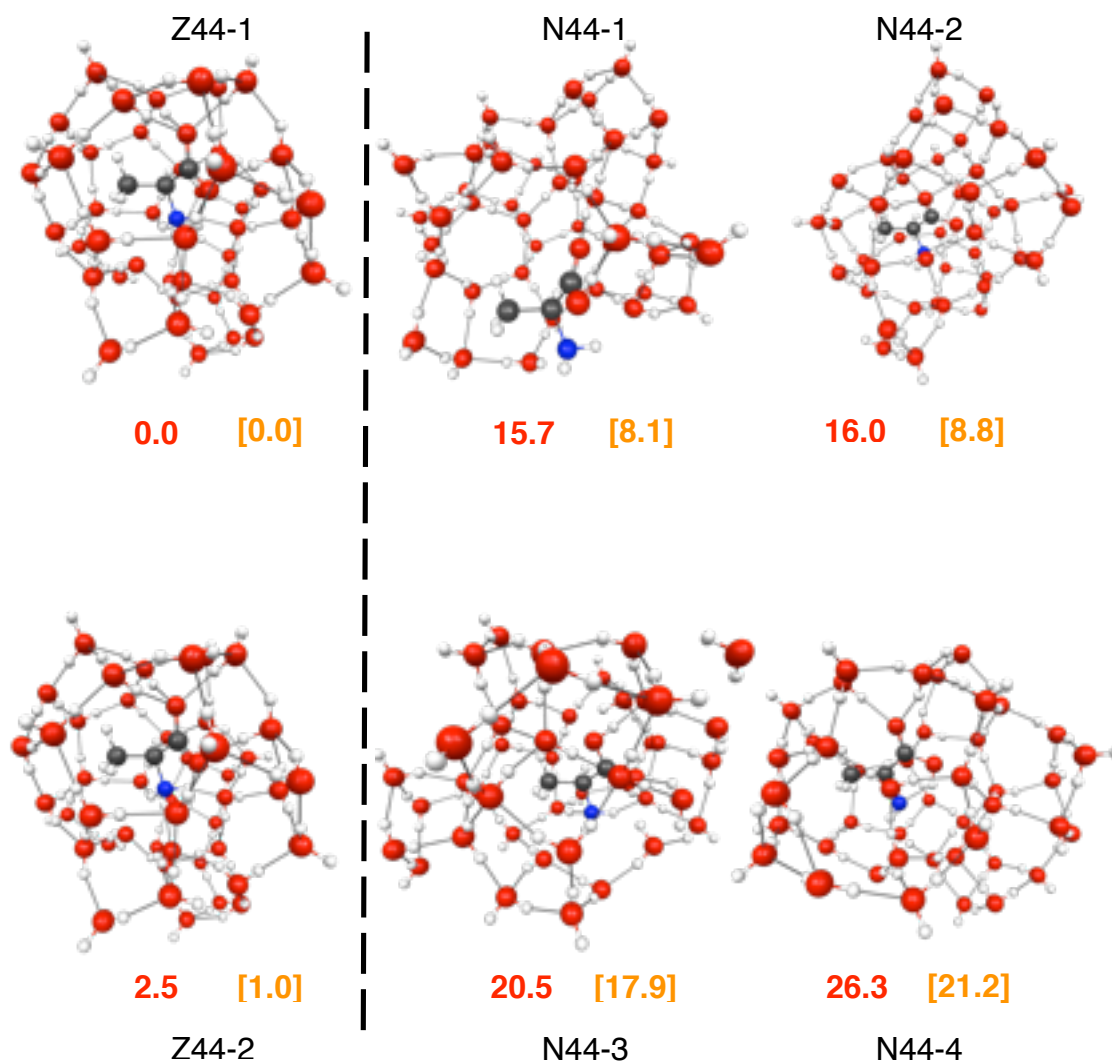
Figure S4. Alanine and 38 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z38-1	N38-1	N38-2
0.0 [0.0]	12.3 [13.2]	18.5 [18.0]

6.6 [19.4]	15.3 [19.4]	28.6 [27.0]
Z38-2	N38-3	N38-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
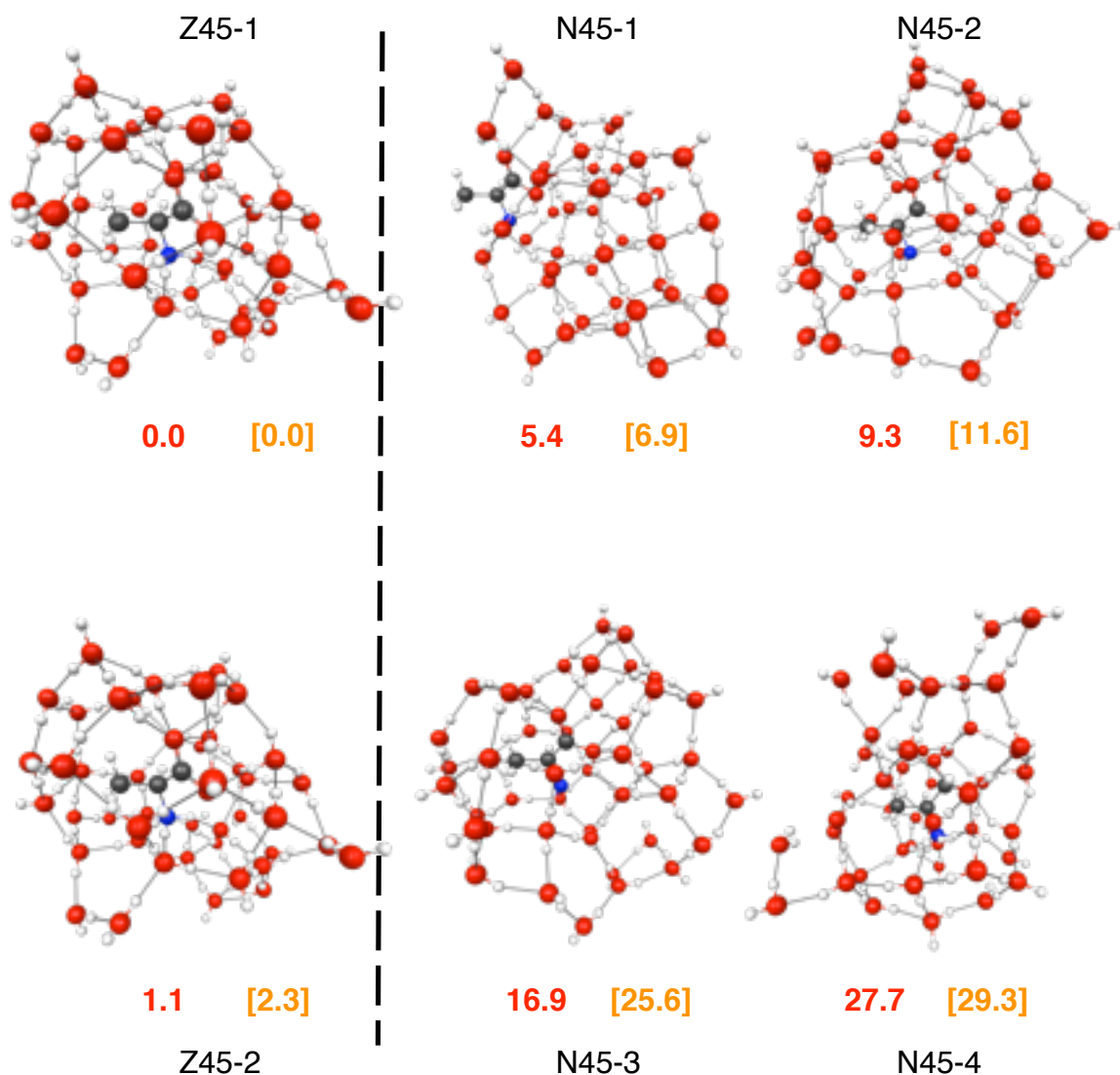PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S5. Alanine and 39 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z39-1   N39-1   N39-2

0.0   [0.0]   6.0   [10.6]   10.21 [13.1]

2.9   [10.8]   11.8   [14.3]   12.8   [14.8]

Z39-2   N39-3   N39-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
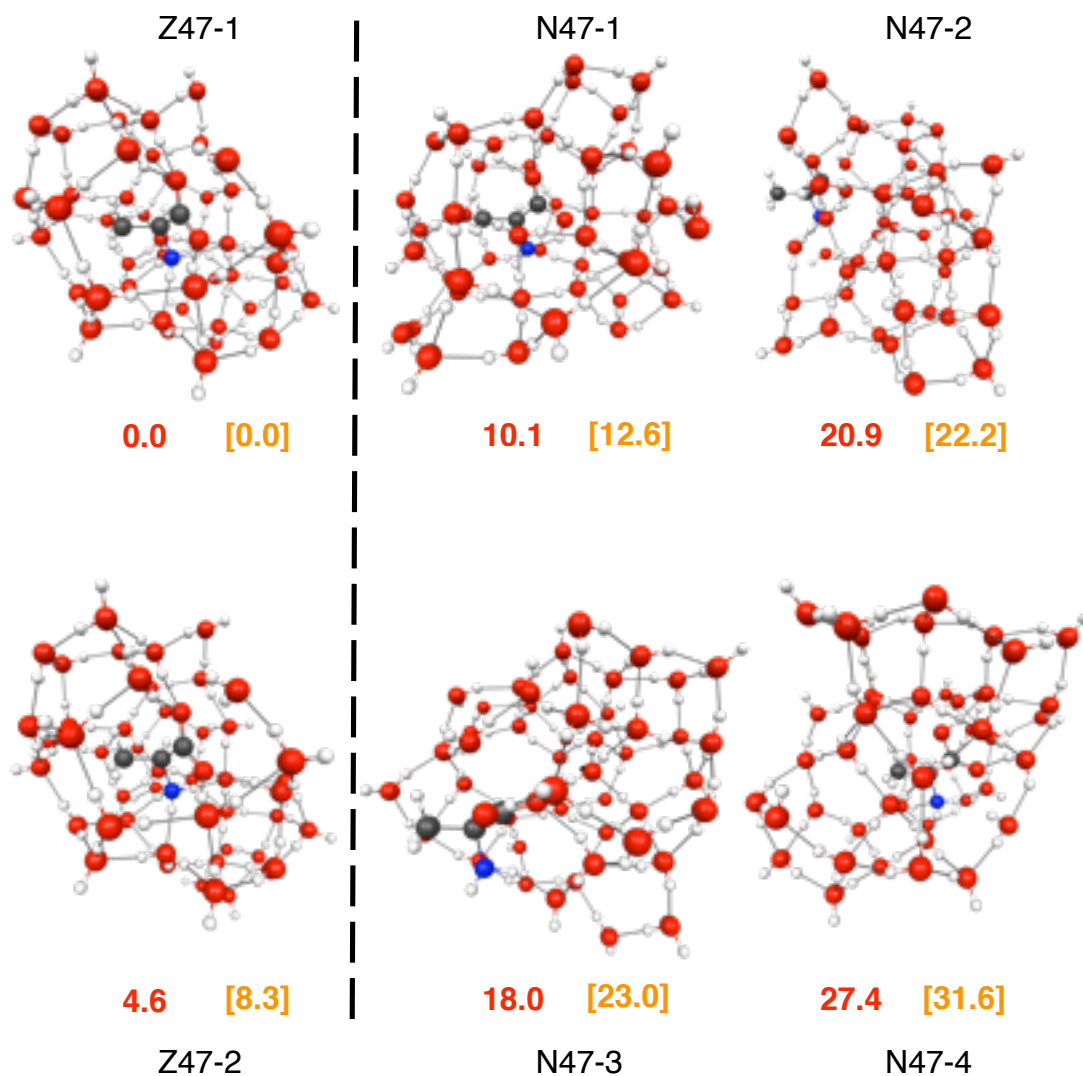PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S6. Alanine and 40 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z40-1   N40-1   N40-2

0.0   [0.0]   15.3   [11.4]   10.7   [12.6]

8.6   [4.2]   27.1   [15.9]   21.2   [26.7]

Z40-2   N40-3   N40-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S7. Alanine and 42 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z42-1

N42-1

N42-2

0.0   [0.0]   **5.8   [6.8]**   **19.0   [18.8]**

5.2   [5.8]   **20.4   [19.7]**   **15.7   [20.6]**

Z42-2

N42-3

N42-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S8. Alanine and 43 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z43-1   N43-1   N43-2

0.0   [0.0]   13.8   [8.8]   10.3   [12.0]

10.3   [12.9]   22.1   [19.1]   17.3   [24.6]

Z43-2   N43-3   N43-4

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S9. Alanine and 44 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z44-1	N44-1	N44-2
0.0 [0.0]	15.7 [8.1]	16.0 [8.8]

Z44-2	N44-3	N44-4
2.5 [1.0]	20.5 [17.9]	26.3 [21.2]

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S10. Alanine and 45 waters. Structures initially from Monte Carlo simula-
tions, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are
shown with representative higher energy structures. Energies (kcal/mol) are relative to
the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT
water) 6-31++G(d,p).

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

Figure S11. Alanine and 47 waters. Structures initially from Monte Carlo simulations, then optimized with MP2(EFP1/DFT)6-31++G(d,p). The Z and N minima are shown with representative higher energy structures. Energies (kcal/mol) are relative to the Z global minimum as identified by PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p).

Z47-1	N47-1	N47-2
0.0 [0.0]	10.1 [12.6]	20.9 [22.2]

| 4.6 [8.3] | 18.0 [23.0] | 27.4 [31.6] |
| Z47-2 | N47-3 | N47-4 |

MP2//MP2 (EFP1/DFT water) 6-31++G(d,p)
PCM+MP2/(EFP1/DFT water)//MP2/(EFP1/DFT water) 6-31++G(d,p)

ACCURATE METHODS FOR LARGE MOLECULAR SYSTEMS

Mark S. Gordon*[‡], Jonathan M. Mullin[‡], Spencer R. Pruitt[‡], Luke B. Roskop[‡], Lyudmila V. Slipchenko[§], and Jerry A. Boatz[&]

[‡] Department of Chemistry, Iowa State University, Ames, IA 50011

[§] Department of Chemistry, Purdue University, West Lafayette, IN 47907

[&] Air Force Research Laboratory, Space and Missile Propulsion Division, AFRL/RZS, 10 E. Saturn Blvd., Edward AFB, CA 93524

Abstract. Three exciting new methods that address the accurate prediction of processes and properties of large molecular systems are discussed. The systematic fragmentation method (SFM) and the fragment molecular orbital (FMO) method both decompose a large molecular system (e.g., protein, liquid, zeolite) into small subunits (fragments) in very different ways that are both designed to retain the high accuracy of the chosen quantum mechanical level of theory while greatly reducing the demands on computational time and resources. Each of these methods is inherently scalable and is therefore eminently capable of taking advantage of massively parallel computer hardware, while retaining the accuracy of the corresponding electronic structure method from which it is derived. The effective fragment potential (EFP) method is a very sophisticated approach for the prediction of non-bonded and intermolecular interactions. Therefore, the EFP method provides a way to further reduce the computational effort

while retaining accuracy, by treating the far field interactions in place of the full electronic structure method. The performance of the methods is demonstrated using applications to several systems, including benzene dimer, small organic species, pieces of the alpha helix, water, and ionic liquids.

1. Introduction

The development of quantum chemistry methods in the 1980s and 1990s primarily focused on performing very accurate calculations on relatively small molecular systems. The desire for accurate calculations on larger molecular species led to several formulations employing more efficient scaling, as well as additivity of basis set improvement and higher levels of electron correlation. With regard to the latter, the Gaussian $G(n)$[1] methods and the Weizmann $W(n)$[2] methods are well known, along with several variants.[3] Because they ultimately rely on the use of very accurate electronic structure methods that scale on the order of n^7, where n measures the size of the system of interest, these approaches are limited fairly small molecular species, with less than 10 heavy (non-hydrogen) atoms.

Simultaneous progress in the development of systematically improving atomic basis sets has also provided a path toward systematic increases in accuracy. It was recognized[4] that basis functions optimized for atomic correlation are also capable of describing molecular correlation effects. Dunning and co-workers, for example, introduced a series of correlation consistent basis set sets[5] based upon these conclusions, capable of accurately treating electron correlation with a compact set of primitive Gaussian functions. These basis sets can be used in a systematic way to obtain results approaching the complete basis set (CBS) limit. However, increasingly large basis sets must be used, and the convergence tends to be slow. Werner has recently introduced a series of F12 basis sets[6] with improved convergence to the CBS limit. The high accuracy

of these basic sets still comes at a significant computational cost, only feasible on relatively small systems.

Chemical phenomena occur in condensed phases as well as in the gas phase, and many methods have been developed to treat the chemical environment[7] and condensed phase phenomena.[8] The desire to study ever larger systems led to combining quantum mechanics (QM) with molecular mechanics (MM). Several such combinations, known as QM/MM methods,[9] have been developed since the initial work of Warshel,[9a] including multi-layer methods such as ONIOM,[10] the Truhlar MCMM methods,[11] and the effective fragment potential method (EFP)[12-27] developed by Gordon and co-workers. The EFP method will be discussed in detail as a means to investigate non-bonded and intermolecular interactions via the automatic generation of a model potential that is derived from first principles.

While hybrid methods have expanded the size of systems that are accessible to computations, the use of classical model potentials for the description of the environment can be a limiting factor, given that the electron density of the MM region and its impact on the QM region is not usually properly accounted for.

Alternative approaches to QM/MM methods are fragmentation methods, in which the system is broken ("fragmented") into smaller pieces, each of which is considered essentially independently by a specified level of electronic structure theory. Fragmentation methods have the advantage that they are nearly fully quantum mechanical in nature, with classical approximations often used for long-range interactions. Several general fragmentation methods have been proposed, including molecular fragmentation with conjugated caps (MFCC),[28] the elongation method,[29] the

molecular tailoring approach (MTA),[30] the fast electron correlation method for molecular clusters developed by Hirata,[31] Truhlar's electrostatically embedded many-body (EE-MB) expansion,[32] multi-centered QM/QM methods,[33] the systematic fragmentation method (SFM),[34-38] and the fragment molecular orbital (FMO) method.[39-45] The latter two methods, the SFM and the FMO methods, will be discussed in detail in this work.

Instead of separating a system into two regions that are described by two very different levels of theory (QM and MM), fragmentation methods that divide a system into many smaller pieces, all of which are described by the same level of QM theory, have been proposed since the 1970s.[46] By approaching a large system in this way, each smaller fragment can be treated using high levels of theory, providing the desired accuracy and an improvement in speed. The earliest attempts[46] constructed a set of fragments from common chemical groups (methyl, amino etc.) and used a selection of these fragments to build larger molecules. More recent fragmentation methods[28-45] begin with the larger molecule of interest and break the system into smaller fragments.

To increase their generality, fragmentation methods should also treat the environment (e.g., the remainder of the entire molecular system, a solvent) around each fragment in some approximate, but realistic manner. When a molecule or a molecular system is fragmented into smaller pieces, each fragment no longer electronically "feels" the remainder of the initial system, unless one devises some way to retain the lost interactions. This issue is addressed in the FMO method[39] by performing each individual fragment calculation in a Coulomb "bath" represented by the electrostatic potential (ESP) of the entire system. Further corrections to the FMO method are achieved by performing fully quantum mechanical two-fragment (dimer) and three-fragment (trimer) calculations.

In the SFM method[34] the effects of other fragments are incorporated by including overlapping fragments in such a manner that the double counting of atoms is accounted for, and non-bonded interactions are captured by employing classical potentials.[12] Accurately capturing non-bonded effects is essential to maintaining kcal/mol accuracy compared to full *ab initio* studies. Both the FMO and SFM methods are discussed in more detail in following sections. In both methods, fully quantum mechanical or fully *ab initio* can refer to any of the common electronic structure methods that are available in most electronic structure packages.

Traditional electronic structure methods, such as Hartree-Fock (HF), second order perturbation theory (MP2), and coupled cluster theory (e.g. CCSD(T)) have rapidly increasing resource requirements (e.g. time, memory, mass storage). For example, the HF, MP2 and CCSD(T) computer time requirements scale as $O(n^4)$, $O(n^5)$ and $O(n^7)$ respectively, where n measures the size of the system, e.g., in terms of the basis set size. Further, CCSD(T) memory requirements scale as $O(n^4)$, while disk requirements are difficult to uniquely define. One approach to addressing the computational scaling issue is to develop highly parallel algorithms. The development of parallel algorithms for electronic structure theory has been an active research area for ~20 years, and considerable progress has been achieved for increasingly complex QM methods.[47] Such efforts may be referred to as fine-grained parallelism, in the sense that each energy or derivative evaluation itself takes advantage of many cores, usually in a distributed manner.[48] In many fragmentation methods each fragment calculation can be performed essentially independently of all the others. This leads to a multi-level parallelism, since the energy of each fragment can be obtained on a separate node (coarse-grained

parallelism), while the fine-grained parallelism can be exploited within each node.[49] If a fragmentation method is implemented to take advantage of this ability, large reductions in required computational resources can be achieved, facilitating calculations on, for example, condensed phases, proteins and surfaces. Fragmentation approaches with multi-level parallelism also expand the capabilities of modest (e.g., single group or departmental) computer systems.

The present work focuses on three methods that have been designed to accurately treat large systems: EFP, SFM, and FMO. As noted above, the semi-classical EFP method has been developed to study non-bonded and intermolecular interactions. Benzene dimer is chosen as a representative example to illustrate the accuracy and efficiency of the EFP method, although several such studies have been carried out,[50] as have EFP molecular dynamics simulations.[51] Both the SFM and FMO methods have been designed to extend fully quantum methods to much larger molecular systems than are commonly accessible, by the development and implementation of judicious approximations. It will be shown that the EFP approach provides an effective means to accurately capture non-bonded interactions within the SFM framework.[39] It will be illustrated that the FMO method can be used to accurately describe a series of water clusters and ionic liquid systems.

2. Effective Fragment Potential (EFP) Method

The generalized effective fragment potential (EFP2) method[12] is a first-principles based model potential for the evaluation of intermolecular forces. This is a modification and extension of the original EFP1 water model[13-17] to general systems. There are five

EFP-EFP interaction terms in the EFP2 model potential, each of which may be thought of as a truncated expansion:

Coulombic (electrostatic), induction (polarization), exchange repulsion, dispersion (Van der Waals), and charge transfer.

$$E = E_{coul} + E_{ind} + E_{exrep} + E_{disp} + E_{ct} \qquad (1)$$

In EFP1, the exchange repulsion, E_{exrep}, and charge transfer, E_{ct}, components are folded into one term that contains fitted parameters, and there is no dispersion contribution. EFP1 has been integrated with HF,[13] DFT,[14] MCSCF,[15] singly excited configuration interaction (CIS),[16] and time-dependent density functional theory (TDDFT).[17] The EFP2-QM interface is still under development.[18]

The main focus in this work is the general EFP2 method. From this point, EFP will imply the EFP2 method. The five terms in the EFP potential may be grouped into long range, $(1/R)^n$ distance dependent, or short range interactions, which decay exponentially. The Coulombic, induction, and dispersion are long-range interactions, whereas the exchange repulsion and charge transfer are short range. EFP has been described in detail in several papers,[12-17] therefore only a brief overview of the terms will be presented below.

The Coulomb portion of the electrostatic interaction, E_{Coul}, is obtained using the Stone distributed multipolar analysis.[19] This expansion is truncated at the octopole term. Atom centers and the bond midpoints are used as expansion points.

Induction (polarization), E_{ind}, arises from the interaction of an induced dipole on one fragment with the permanent dipole on another fragment, expressed in terms of the dipole polarizability. Truncating at the first (dipole) term in the polarizability expansion

is viable, since the molecular polarizability tensor is expressed as a tensor sum of localized molecular orbital[20] (LMO) polarizabilities. Therefore, the number of bonds and lone pairs in the system is equal to the number of polarizability points. This induction term is iterated to self-consistency, so it is able to capture some many body effects.

Because the Coulomb and induction terms discussed above, as well as the dispersion interaction, are treated primarily by classical approximations, the shorter range interactions that occur when quantum mechanical charge densities begin to overlap are not correctly captured. Therefore, each term is multiplied by a damping (screening) expression. The relative merits of several approaches to damping have recently been analyzed and discussed extensively.[22] Classical Coulombic interactions become too repulsive at short range, and must be moderated by a screening term, as discussed in several previous papers.[21,22] Conversely, the induction interaction becomes too attractive in the short-range regime, so a damping term is needed here as well. The unphysical behavior is avoided by augmenting the electrostatic multipoles with exponential damping functions of the form:

$$f_{\text{damp}} = 1 - \exp(-\alpha R) \tag{2}$$

where parameters α are determined at each multipole expansion point by fitting the multipole damped potential to reproduce the Hartree–Fock potential. Damping terms in the electrostatic energy are derived explicitly from the damped potential and the charge density. The damping procedure can be extended to higher-order electrostatic terms; that is, the charge– dipole, dipole–dipole, etc., interactions, and this is recommended.[21] Damping is also applied to the induction and dispersion energies.[21,22] For induction, both exponential damping, as in Eq. (2) and Gaussian damping are effective, but the Gaussian

damping seems to be more generally applicable and is therefore recommended.

The exchange repulsion interaction between two fragments is derived as an expansion in the intermolecular overlap.[23] When this overlap expansion is expressed in terms of frozen LMOs on each fragment, the expansion can reliably be truncated at the quadratic term. This term does require each EFP to carry a basis set. Since the same basis set is used to generate the multipoles and the molecular polarizability tensor, EFP calculations are basis set dependent. The smallest recommended basis set is 6-31++G(d,p)[52] The dependence of the computational cost of an EFP calculation on the basis set appears primarily in the initial generation of the EFP. Therefore, one can employ much larger basis sets with minimal cost. The tests presented below on the SFM method use the 6-311++G(3df,2p)[53] basis set. Since the basis set is used only to calculate overlap integrals, the computation is very fast and quite large basis sets are realistic.

Dispersion interactions are often expressed by an inverse R expansion,

$$E_{disp} = \sum_n C_n R^{-n} \qquad (3)$$

where the coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities integrated over the entire frequency range.[24] The first term in the expansion, n=6, corresponds to the induced dipole-induced dipole (Van der Waals) interactions. In the EFP2 method, this term is evaluated using the time-dependent HF method. In addition the contribution of the n=8 term is estimated. The C_6 coefficients are derived in terms of interactions between pairs of LMOs, one each on two interacting molecular species, or EFPs. Because the dispersion interaction should decrease to zero at short range, each dispersion term is multiplied by a damping function. Tang-Toennies damping[25] is frequently used to damp dispersion. However, a new approach that is based

on the overlap integrals between interacting fragments[22] is free of fitted parameters and appears to be generally applicable. In future EFP applications, the overlap-based dispersion damping is recommended.

The charge transfer interaction is derived using a supermolecule approach, in which the occupied valence molecular orbitals on one fragment are allowed to interact with the virtual orbitals on another fragment. This interaction term leads to significant energy lowering in *ab initio* calculations on ionic or highly polar species when incomplete basis sets are employed. An approximate formula[26] for the charge transfer interaction in the EFP2 method was derived and implemented using a second order perturbative treatment of the intermolecular interactions for a pair of molecules at the Hartree–Fock level of theory. This approximate formula is expressed in terms of the canonical orbitals from a Hartree–Fock calculation of the isolated molecules and uses a multipolar expansion (through quadrupoles) of the molecular electrostatic potentials. Orthonormality is enforced between the virtual orbitals of the other molecule and all of the orbitals of the considered molecule, so that the charge transfer is not contaminated with induction. This approximate formula has been implemented in the EFP method and gives charge transfer energies comparable to those obtained directly from Hartree Fock calculations.[26] The analytic gradients of the charge transfer energy were also derived and implemented, enabling efficient geometry optimization.[27]

It is useful to consider the relative costs of the five EFP interaction terms. Based on relatively small molecules and taking the cost of the Coulomb and dispersion terms to be one unit, the induction interaction would cost approximately two units, exchange repulsion would cost about five units, and charge transfer would cost ~50 units. For

larger molecules, the relative costs of exchange repulsion and charge transfer will decrease since they will scale linearly in the large molecule limit. As always in computational chemistry there is a trade off between computational cost and accuracy.

While the EFP model is currently a rigid body model potential, analytic gradients for all terms have been derived and implemented, so full intermolecular geometry optimizations, Monte Carlo and molecular dynamics simulations[50,51] can be performed. Because the method involves no empirically fitted parameters, an EFP for any system can be generated by a "makefp" run in the GAMESS[54] suite of programs. The EFP potential generated by the makefp run includes (i) multipoles (produced by the Stone Distributed Multipolar Analysis) that are used in calculations of Coulomb and polarization terms; (ii) static polarizability tensors centered at LMOs obtained from CPHF calculations, which are used for calculations of the polarization energy and gradients; (iii) dynamic polarizability tensors centered on the LMOs that are generated generated by TDHF calculations and used for calculations of dispersion; (iv) the Fock matrix, basis set, and localized orbitals for the exchange-repulsion term, and (v) canonical orbitals for the charge-transfer term. This automatic generation makes possible the use of the EFP method for treating intermolecular and non-bonded interactions in fragmentation methods such as the SFM.

2.1 The EFP Method as a Model for Non-Bonded Interactions

Benzene dimer is used here to illustrate the accuracy of EFP-EFP non-bonded interactions, with a focus on the $\pi-\pi$ interactions between two benzene rings. These $\pi-\pi$ interactions are largely driven by dispersion and are therefore difficult to account for accurately by most *ab initio* electronic structure methods. Previous theoretical and

experimental studies suggest that there are two minima on the benzene dimer potential energy surface[55-62] the perpendicular T-shaped and parallel-slipped configurations, as shown in Figure 1. A sandwich structure with two parallel benzene rings, also shown in Figure 1, is a saddle point that connects two equivalent parallel slipped structures. Sherrill and co-workers calculated potential energy curves for these three structures[56,57] using second order Møller-Plesset perturbation theory (MP2) and coupled-cluster theory with single, double and perturbative triple excitations [CCSD(T)][63] A variety of augmented correlation consistent basis sets[8] were used with both the MP2 and CCSD(T) levels of theory. Additionally, they employed symmetry-adapted perturbation theory (SAPT)[64] to decompose the benzene $\pi - \pi$ interaction energy into electrostatic, dispersion, induction and exchange-repulsion components of the total interaction energy. The binding energies, equilibrium separations and SAPT energy decomposition results from their work compare well with similar results obtained using the EFP method, as illustrated below.[21a]

For this work, the EFP for benzene was constructed with the 6-311++G(3df,2p) basis set,[53] using the MP2/aug-cc-pVTZ[64] benzene monomer geometry taken from Ref. 57. The multipoles for benzene were generated using a numerical distributed multipolar analysis (DMA).[19] The numerical DMA scheme is employed due to the instability and basis set dependence of the standard analytic DMA scheme, as well as the need for diffuse functions to properly describe the exchange-repulsion interactions within the EFP framework. Higher order (up to quadrupoles) damping terms were also used to provide an accurate description of charge penetration through screening of the potentials.[21]

The EFP binding energies and corresponding inter-ring distances for the three

benzene dimer structures are in good agreement with the analogous *ab initio* values obtained by Sherrill and co-workers (See Table 1). Relative to the full CCSD(T)/aug-cc-pVQZ binding energies, the EFP method over-binds the sandwich dimer by 0.4 kcal/mol and under-binds the T-shaped structure by 0.1 kcal/mol, while the equilibrium intermolecular separations are overestimated by approximately 0.1–0.2 Å. In comparison, MP2 with the same basis set overestimates the binding energies by 1.7 kcal/mol and 0.9 kcal/mol for the sandwich and T-shaped dimers, respectively, and underestimates the equilibrium distances by approximately 0.1–0.2 Å. In fact, the MP2 binding energies become successively worse compared with those predicted by CCSD(T) as the basis set is improved. The EFP and CCSD(T) predicted binding energies and structures are in reasonable agreement with each other, whereas the agreement between MP2 and CCSD(T) is not as good. Table 1 summarizes the MP2, CCSD(T) and EFP total interaction energies of all three benzene dimer structures. A comparison of the total interaction energy decompositions obtained using both SAPT and the EFP method shows good agreement for the sandwich and T-shaped isomers (See Figures 2 and 3). Specifically, the error in the EFP method compared to SAPT for the dispersion, exchange-repulsion and polarization interactions is in the range 0.2-0.5 kcal/mol for these two isomers.[21a]

Highly accurate methods involve very demanding scaling of computational resources, such as time, memory and disk. For instance, a single point energy calculation in the 6-311++G(3df, 2p) basis set (660 basis functions) by MP2 requires 142 minutes of CPU time on one IBM Power5 processor, whereas the analogous EFP calculation requires only 0.4 seconds. The corresponding CCSD(T) calculation would be much more

resource demanding than MP2. Taking into account the relatively good agreement of the EFP method results with the CCSD(T)/aug-cc-pVQZ results described above, this significant reduction in total computation time comes with a minimal loss of accuracy.

3. The Systematic Fragmentation Method (SFM)

The systematic fragmentation method (SFM) is designed to permit a large molecular system, such as a protein, a polymer, or a surface, to be fragmented into smaller pieces in such a way that retains the accuracy of a full *ab initio* calculation at the same level of theory, while significantly decreasing the computational expense. By treating the smaller sub-systems with accurate levels of theory, the total energy and properties of the full system are obtained through addition and subtraction of the contributions from the overlapping sub-systems or "groups." Many body effects are accounted for including the nearest neighbor of each group. Non-bonded interactions between groups are also accounted for. In the original formulation[34] these non-bonded interactions were obtained using a classical electrostatic potential. Recently, as illustrated in the following paragraphs, this non-bonded description has been improved through the use of the EFP method, providing a more accurate representation of the non-bonded interactions.[38]

Within the context of the SFM, a molecule can be thought of as a collection of functional groups. For example, ethanol contains three functional groups (CH_3, CH_2, and OH) according to the SFM prescription. To fragment the system into functional groups, single bonds are broken. This process splits a pair of bonding electrons; each of these electrons is assigned to one of the two resulting fragments. In order to avoid the resulting radical species, a hydrogen atom is used to "cap" the dangling bonds that are created by

the fragmentation. The capping hydrogen points in the direction of the broken bond at a chemically reasonable distance. By design, double or triple bonds are not broken, keeping the relevant atoms as a part of one functional group. For example, ethanal would contain two functional groups (CH_3 and CHO), keeping the carbon and oxygen atoms of the carbonyl together in one group. After the addition of the hydrogen caps, the ethanol groups would be CH_4, CH_4, and H_2O, and the ethanal groups would be CH_4 and CH_2O.

To gain a more quantitative understanding of the SFM, consider the general example of an acyclic molecule M containing K functional groups G_i:

$$M = G_1 G_2 G_3 ... G_k . \qquad (4)$$

Each group G_i is connected by single bonds to adjacent groups G_{i-1} and G_{i+1}. In order to separate the functional groups of M into smaller fragments, one can imagine breaking the $G_{i-1} - G_i$ single bond, then capping each new terminal atom with a hydrogen atom. This produces two new, smaller species,

$$M_1 = G_1 G_2 G_3 ... G_{i-1} H_{i-1} \qquad (5)$$

$$M_2 = H_i G_i G_{i+1} ... G_k . \qquad (6)$$

The internal geometries of M_1 and M_2 are preserved, except for the hydrogen atoms that have been used to cap the missing bond vector. The total energy can then be written, without approximation, as:

$$E(M) = E(M_1) + E(M_2) + dE_1 , \qquad (7)$$

where dE_1 is the correction for the differential change in the energy caused by breaking a bond and adding two hydrogen caps. This process can be repeated, since bonds can be broken at any point along the chain, decomposing the full system into many smaller fragments. As the separation between the bond breaks is increased, the accuracy of the

SFM will increase, since the larger fragments will give a more accurate description of the full system. The separation between broken bonds can be described as different "levels" of the SFM.

The SFM levels are defined as follows.[34]

Consider the molecule M:

$$M = G_1G_2G_3G_4G_5G_6G_7G_8 \qquad (8)$$

In the Level 1 SFM, two bonds separated by just one functional group are sequentially broken. The fragments initially created would, for example, be as follows:

$$M \approx G_1G_2 + G_2G_3G_4G_5G_6G_7G_8 - G_2 \qquad (9)$$

The G_2 fragment is subtracted off to conserve the number of atoms. Subsequently, this process is repeated exhaustively on the $G_2G_3G_4G_5G_6G_7G_8$ fragment until no fragment larger than 2 functional groups remains. In the end, the energy of M can be approximately decomposed into the simple sum of fragment energies for level 1 as follows:

$$\begin{aligned} E_{\text{level 1}}^{\text{bonded}}(M) = {}& E(G_1G_2) + E(G_2G_3) + E(G_3G_4) + E(G_4G_5) + E(G_5G_6) + E(G_6G_7) + E(G_7G_8) \\ & - E(G_2) \quad - E(G_3) \quad - E(G_4) \quad - E(G_5) \quad - E(G_6) \quad - E(G_7) \end{aligned}$$

$$(10)$$

Similarly in the Level 2 SFM, bonds separated by two functional groups are sequentially broken with the energy of M being decomposed into the following expression:

$$\begin{aligned} E_{\text{level 2}}^{\text{bonded}}(M) = {}& E(G_1G_2G_3) + E(G_2G_3G_4) + E(G_3G_4G_5) + E(G_4G_5G_6) + E(G_5G_6G_7) + E(G_6G_7G_8) \\ & - E(G_2G_3) \quad - E(G_3G_4) \quad - E(G_4G_5) \quad - E(G_5G_6) \quad - E(G_6G_7) \end{aligned}$$

$$(11)$$

In the level 3 SFM, bonds separated by three functional groups are sequentially broken with the energy of M being decomposed into the following expression:

$$\begin{aligned} E_{\text{level 3}}^{\text{bonded}}(M) = {}& E(G_1G_2G_3G_4) + E(G_2G_3G_4G_5) + E(G_3G_4G_5G_6) + E(G_4G_5G_6G_7) + E(G_5G_6G_7G_8) \\ & - E(G_2G_3G_4) \quad - E(G_3G_4G_5) \quad - E(G_4G_5G_6) \quad - E(G_5G_6G_7) \end{aligned}$$

(12)

It is important to note that in the limit of SFM, that is for level n, where n is the number of groups in the system, one would be left with the un-fragmented system. So the higher the SFM level employed, the larger the fragments and the closer one should get to the energy of the exact un-fragmented system.

There are some limitations of the SFM. First, as noted earlier, the SFM is unable to fragment conjugation in delocalized molecular systems. The second, less obvious, limitation is that the SFM is unable to fragment six member rings using level 3 since the capping hydrogens would approach each other too closely and would therefore cause unphysical repulsive interactions. To avoid this, the ring must remain intact and is considered to be a functional group itself. Similarly, five member rings can only be fragmented at level 1; four and three member rings cannot be fragmented at all. These exceptions are referred to as the ring repair rule.

3.1 Non-bonded interactions

The simplest approach to obtain the energy of the system of interest would be to calculate the energies of the individual hydrogen-capped fragments and sum them accordingly. The result obtained from this procedure would differ greatly from the analogous calculation on the full molecular system. This is because the (non-bonded) interactions among the separated fragments are unaccounted for. These non-bonded interactions are naturally incorporated into the full *ab initio* calculation. The non-bonded interactions are modeled within the SFM framework by using a modified many body expansion;[37] this expansion relies on the assumption that bonded interactions are much stronger than non-bonded ones.

3.2 Two-body interactions

The interaction energy between two functional groups G_1 and G_2 is given by

$$E_{nb}^{(1,1)}[G_1;G_2] = E(G_1G_2) - E(G_1) - E(G_2),\qquad\qquad(13)$$

where $E(G_1G_2)$ is the super-molecular energy of the two separated functional groups (placed in their positions in the original full molecule M) and $E(G_1), E(G_2)$ are the corresponding (one-body) fragment energies. The total two-body non-bonded energy of the system contains the energies of all possible pairs of functional groups that are not described by the fragmentation of the bonded system in the definition of *M*. That is, all pairs of groups G_1, G_2 that are not contained in any one fragment.

3.3 Three-body interactions

The mutual interaction of three functional groups G_1, G_2 and G_3 is assumed to be negligible unless any two of the groups are bonded to each other. For example, if G_3 is bonded directly to G_2 then the three-body interaction energy would be:

$$\begin{aligned}E_{nb}^{(1,2)}[G_1;G_2,G_3] &= E(G_1G_2G_3) - E(G_1) - E(G_2G_3)\\ &\quad - E_{nb}^{(1,1)}[G_1;G_2] - E_{nb}^{(1,1)}[G_1;G_3]\end{aligned}\qquad(14)$$

In other words, the three-body energy is simply the super-molecular energy, $E(G_1G_2G_3)$, minus the one-body energies $E(G_1), E(G_2G_3)$ and minus the two-body energies, $E_{nb}^{(1,1)}[G_1;G_2], E_{nb}^{(1,1)}[G_1;G_3]$. The total three-body energy consists of all combinations containing any group (G_1) with any two bonded functional groups (G_2 or G_3), so long as G_1 is itself not present in any bonded fragment with G_2 or G_3. This general trend can be extended to four body interactions and beyond; however for the purposes of this work only three-body terms will be treated. Note that to employ the SFM method, one only

needs to specify the desired Level. The fragmentation then follows without further specification.

The total SFM energy of a system is simply the addition of the bonded and non-bonded energies,

$$E_{SFM}^{\text{total}} = E^{\text{bonded}} + E^{\text{non-bonded}}, \qquad (15)$$

where $E^{\text{non-bonded}}$ includes all terms up to n^{th} order from the modified many-body approximation. For example, calculations employing 3rd order many body non-bonded energies would include the 2nd order many body non-bonded energies as well.

3.4 SFM and EFP

SFM molecular energy calculations corresponding to bonded level 3 including many body non-bonded interactions apparently provide, on average, the best balance between accuracy and computational effort.[35] Although the non-bonded approximation is important for reliability, it also hinders computational performance by significantly increasing the number of *ab initio* terms. For example, moderately sized proteins (~3500 residues) have on the order of 10^6 non-bonded interactions. Because there are so many non-bonded terms, these terms can dominate the calculation. It is therefore advantageous to employ approximate methods for those non-bonded interactions that are sufficiently distant that classical approximations might be valid. The simplest approach, using just electrostatic interactions, were used in the original SFM implementation.[35] A more sophisticated approach, using effective fragment potentials (EFP), is described here[38]. Compared to electrostatics, intermediate range (2.7-4.5Å) EFP interaction energies agree better with *ab initio* methods. This increases the number of non-bonded terms that can be calculated with model potentials.

The determination of whether a non-bonded term is treated with EFP or *ab initio* methods is based on a user-defined cutoff related to the nearest atom-atom distance between fragments. The short-range (< 2.7Å) non-bonded terms use *ab initio* methods, while long-range (≥2.7Å) ones use EFP. The original electrostatic approach[38] used a cutoff of 4.5Å. This shortened EFP cutoff comparatively reduces the number of *ab initio* non-bonded terms, thereby decreasing the computational expense.

Previously, Collins and Deev tested the SFM by calculating the isomerization energies of a set of organic molecules (12-44 heavy atoms) obtained from the Cambridge Structural Database.[65] A subset of this set of isomerization energies is examined here. The energies that are obtained by employing the EFP/*ab initio* non-bonded approach are compared with both the fully *ab initio* energies (no SFM) and to the SFM in which all non-bonded terms are calculated with the same *ab initio* method that is used for the bonded terms. The *ab initio* calculations here employ both the Hartree-Fock (HF) and second order perturbation theory (MP2) levels with the 6-31G(d,p) basis set. Additional SFM tests are presented for a small set of alpha helices using the 6-31++G(d,p) basis set. The larger 6-311++G(3df,2p) basis set is employed for creating all EFPs used for non-bonded interactions, since this basis set has been shown to produce reliable results and since the EFP basis set dependence does not significantly affect the computational cost relative to *ab initio* calculations. All of the SFM calculations presented here correspond to bonding level 3, including up to 3rd order many body non-bonded interactions. All calculations are performed with the GAMESS[54] electronic structure code.

Given in Table 2 are the errors in the isomerization energies. The corresponding structures are depicted in Scheme 1. It is evident that the two methods for treating the

SFM non-bonded energy (EFP/*ab initio* and *ab initio* only) are in reasonable agreement with the fully *ab initio* (non SFM) energies, as the mean absolute error (MAE) in all cases is no more than 2.5 kcal/mol. Addition of the 3rd order non-bonded terms does not result in any improvement to the MAE. Interestingly, the MAE for the combined EFP/*ab intio* approach for the non-bonded terms are slightly smaller (~0.1-0.5 kcal/mol) than those obtained when the non-bonded terms are evaluated with the *ab initio* method (HF or MP2). For the 21 molecules of interest here, as also noted by Collins and Deev[35], no improvement in the net CPU time is observed since the molecules themselves are small. Improvements in CPU timings are observed for larger molecular systems (>100 atoms), as discussed below.

SFM isomer energies for the larger model alpha helices (ranging from 125-170 atoms) are shown in Table 3, with the corresponding structures presented in Scheme 2. For these systems, adding the higher order non-bonded terms does improve the SFM performance. The MAE improves by ~1 kcal/mol when the 3rd order non-bonded terms are included. Here again, the SFM errors obtained when using the EFP/*ab initio* non-bonded energies are similar (~1 kcal/mol *smaller*) to those obtained using only *ab inito* non-bonded terms. Table 4 compares the CPU times for using the EFP method for non-bonded terms with those required for the *ab initio*-only SFM. The time needed to generate the EFP terms is also listed. This time becomes significant when the 3rd order many body terms are included. Further, since the EFP generation requires only calculations at the Hartree-Fock level of theory, the contribution of the EFP generation to the overall computation time will greatly decrease in importance when more accurate electronic structure methods are used.

Nonetheless, as shown in Table 4, employing EFP to treat a portion of the SFM 3rd order non-bonded terms results in an overall decrease in CPU time by roughly a factor of two. Including only the 2nd order non-bonded EFP/*ab initio* terms gives energies in good agreement with the full un-fragmented energies (Table 3: MAE = 2.6 kcal/mol), but the gain in computational efficiency is small, ~5-15% less CPU time. The advantage of using the EFP/*ab initio* approach is clearly seen in Table 5, where the number of non-bonded terms that must be computed *ab initio* is compared for the EFP/*ab initio* and the electrostatic/*ab initio* methods. Since the EFP method is more effective at capturing interaction energies than electrostatics at close range, the EFP non-bonded cutoff can be set to the shorter distance of 2.7Å, instead of 4.5Å. This shorter cutoff value reduces the number of expensive *ab initio* non-bonded terms by up to 85-90%, while still retaining good accuracy. This increase in efficiency will be especially important when high levels of theory, such as MP2 or coupled cluster methods, are employed to treat large molecular systems. A major advantage of the SFM (and other fragment-like methods) is that it enables very accurate calculations on large molecular systems that would otherwise be impossible. As noted above, since the EFP generation requires only Hartree-Fock level calculations, the contribution of the EFP generation to the overall computation time will greatly decrease in importance when more accurate electronic structure methods are used.

4. The Fragment Molecular Orbital (FMO) Method

The FMO method[39-45] relies upon the assumption that electron exchange and charge transfer are largely local phenomena in chemical systems. By breaking a system into fragments and treating the long-range interactions in a system using only a Coulomb operator, there are significant reductions in computational expense. In addition to this

initial reduction in computational cost, the FMO method is further enhanced with the generalized distributed data interface (GDDI).[49] The GDDI uses a two-level parallelization scheme, assigning individual fragment calculations to different groups, each group performing its fragment calculation in parallel. The FMO method has also been interfaced with the polarizable continuum model (PCM)[66] and the effective fragment potential (EFP)[84] for the inclusion of solvent effects. There is also a multi-layer FMO (MFMO) implementation[67] that allows for the use of different wavefunction types for different fragments. The combination of the long-range approximations to the system and the GDDI parallelization helps to facilitate the treatment of large molecular systems.[45,68]

Creating fragments in the FMO method involves breaking bonds electrostatically, assigning two electrons of a covalent bond to one fragment and none to the other, with the fragment choice relying on the chemical intuition of the user. To avoid the charged species created by such a fragmentation scheme, a proton from the electron deficient fragment is reassigned to the electron rich species, creating two neutral fragments (indicated by the "1" and "5" in Figure 4). The "1" and "5" in the figure *both* carry sp^3 hybrid orbitals, to maintain the carbon character. The individual fragment (monomer) calculations are performed in the presence of a Coulomb "bath" that represents the electrostatic potential (ESP) of the system (Figure 5). As described below, the Coulomb bath is treated by a variety of approximate methods that depend on the distance that separates monomers, dimers, trimers, etc. Significant improvements[69,70] to this description of the FMO method are obtained by including many-body effects. Two-body effects are incorporated by explicitly including all pairs of fragments (monomers). These

pairs are called dimers, and the FMO method that includes them is called FMO2. Likewise, three-body effects are embodied in the FMO3 version of the method, in which all trimers are explicitly included. In FMO2 (FMO3) all dimers (trimers) are treated with the chosen level of electronic structure theory.

To calculate the energy of a system within the FMO method, first the initial electron density distribution is calculated for each monomer in the Coulomb bath of the system. The monomer Fock operators are constructed and the energy of each monomer is calculated. Each of the monomer energies is iterated to self-consistency in this manner, leading to the convergence of the ESP.

The total energy of a chemical system, within the FMO approximation, can be written as;

$$E = \sum_{I}^{N} E_I + \sum_{I>J}^{N} (E_{IJ} - E_I - E_J)$$

$$+ \sum_{I>J>K}^{N} \{(E_{IJK} - E_I - E_J - E_K) - (E_{IJ} - E_I - E_J) \tag{10}$$

$$-(E_{JK} - E_J - E_K) - (E_{KI} - E_K - E_I)\} + ...$$

where monomer (*I*), dimer (*IJ*) and trimer (*IJK*) energies are obtained using the standard SCF method. Despite the seeming simplicity of Eq. (10), the FMO method encapsulates the deeper ideas of properly handing many-body effects, as clarified in the diagrammatic treatment[69] and the Green's function formalism.[73] This is a very important distinction between the FMO and other methods. The Fock equation;

$$\tilde{\mathbf{F}}^x \mathbf{C}^x = \mathbf{S}^x \mathbf{C}^x \tilde{\varepsilon}^x \quad x = I, IJ, IJK \tag{11}$$

$$\tilde{\mathbf{F}}^x = \tilde{\mathbf{H}}^x + \mathbf{G}^x \tag{12}$$

is modified from the standard form with the addition of a term, $V_{\mu v}^{x}$, that represents the ESP to the one-electron Hamiltonian $\tilde{\mathbf{H}}^{x}$.

$$\tilde{H}_{\mu v}^{x} = H_{\mu v}^{x} + V_{\mu v}^{x} + B\sum_{i}\left\langle\mu\middle|\varphi_{i}^{h}\right\rangle\left\langle\varphi_{i}^{h}\middle|v\right\rangle \tag{13}$$

The modified Hamiltonian also contains the projection operator, $B\sum_{i}\left\langle\mu\middle|\varphi_{i}^{h}\right\rangle\left\langle\varphi_{i}^{h}\middle|v\right\rangle$, needed for division of basis functions along the fractioned bonds, where B is a constant chosen to be sufficiently large to remove the corresponding orbitals out of variational space (normally $B=10^{6}$ a.u.).

The ESP of the system takes the form;

$$V_{\mu v}^{x} = \sum_{K(\neq x)}(u_{\mu v}^{K} + v_{\mu v}^{K}) \tag{14}$$

$$u_{\mu v}^{K} = \sum_{A\in K}\left\langle\mu\middle|(-Z_{A}/|\mathbf{r}-\mathbf{r_{A}}|)\middle|v\right\rangle \tag{15}$$

$$v_{\mu v}^{K} = \sum_{\lambda\sigma\in K}D_{\lambda\sigma}^{K}(\mu v\,|\,\lambda\sigma) \tag{16}$$

where the first term $u_{\mu v}^{K}$ is the nuclear attraction contribution and the second term $v_{\mu v}^{K}$ is the two-electron contribution, both of which are calculated for each of the surrounding monomers K with electron density D^{K}.

4.1 FMO Approximations

The formulation of the energy described above has limitations[44,70], such as the increasing cost of the two-electron term in the ESP. To reduce this cost, different approximations can be used to treat the ESP by creating a cut-off value R_{app}. Outside this cut-off the two-electron terms of the ESP can be treated in a more approximate way. However, the foregoing energy formulation loses some accuracy with such

approximations, because the balance among the approximations in different FMO terms may be lost. For example, if there are three monomers I, J and L with some distance based approximation (R_{app}) applied, and the relative distances are as illustrated in Figure 6, then the electrostatic interaction of monomers I and L would be treated using the approximation, while the interaction of monomers J and L would be treated with the full ESP. However, there would be an interaction of dimer IJ with monomer L without any approximations, (because L is close to IJ and J, but far from I). This causes a loss of balance among some of the dimer energy terms in the expression

$$\sum_{I>J}^{N} (E_{IJ} - E_I - E_J) \tag{17}$$

for those dimers IJ in which some ESP contributions (i.e., those for fragment L) included in E_I are treated using the approximation, but in others they are not. There can be a great many dimer contributions to the energy in a single calculation, causing significant loss of accuracy in the energy of the full system.

The issue described above requires a reformulation of the energy that is equivalent to Eq. (10), but more accurate if approximations to the ESP are used[44]. For FMO2,

$$E_{mol} = \sum_I E_I' + \sum_{I>J} (E_{IJ}' - E_I' - E_J') + \sum_{I>J} \{Tr(\mathbf{D}^{IJ}\mathbf{V}^{IJ}) - Tr(\mathbf{D}^I\mathbf{V}^I) - (\mathbf{D}^J\mathbf{V}^J)\}$$

$$= \sum_I E_I' + \sum_{I>J} (E_{IJ}' - E_I' - E_J') + \sum_{I>J} Tr(\Delta\mathbf{D}^{IJ}\mathbf{V}^{IJ}) \tag{18}$$

A similar expression has been derived for FMO3.[78]

The new energy terms E_x' are defined as the internal energies of the monomers and dimers with the ESP contributions subtracted out;

$$E'_x = E_x - Tr(\mathbf{D}^x \mathbf{V}^x) \qquad x = I, J, IJ \tag{19}$$

This is accomplished by contracting $\mathbf{V}^x$ with the electron density $\mathbf{D}^x$. $\Delta\mathbf{D}^x$ is the difference density matrix, defined as

$$\Delta\mathbf{D}^{IJ} = \mathbf{D}^{IJ} - \mathbf{D}^I \oplus \mathbf{D}^J = \begin{pmatrix} \mathbf{d}^{II} & \mathbf{d}^{IJ} \\ \mathbf{d}^{JI} & \mathbf{d}^{JJ} \end{pmatrix} - \begin{pmatrix} \mathbf{d}^I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{d}^J \end{pmatrix} \tag{20}$$

where $\mathbf{d}^{II}$, $\mathbf{d}^{IJ}$, $\mathbf{d}^{JI}$ and $\mathbf{d}^{JJ}$ are blocks of $\mathbf{D}^{IJ}$, and $\mathbf{d}^I$ ($\mathbf{d}^J$) is simply equal to $\mathbf{D}^I$ ($\mathbf{D}^J$). This formulation makes it possible to calculate the total energy explicitly from only the dimer ESP $\mathbf{V}^{IJ}$. By subtracting the monomer and dimer ESPs in the energy expression, approximations can be applied to the monomers and dimers separately. The dimer ESP then directly contributes to the total energy, while the monomer ESP determines the monomer electron densities, only contributing to the total energy indirectly.

Two different levels of approximation are currently used in the FMO method, enabled by equation (18). For intermediate distances the Mulliken approximation[71] to the two-electron integrals is used. Equation (16) can then be rewritten as;

$$v_{\mu v}^K \cong \sum_{\lambda \in K} (\mathbf{D}^K \mathbf{S}^K)_{\lambda\lambda} (\mu v \mid \lambda\lambda) \tag{21}$$

This approximation reduces the computational cost of the two-electron integrals by a factor of N_B (number of basis functions).

The fractional point charge approximation, using the Mulliken atomic populations of the monomers, is used for long distances. The two-electron term of equation (16) is then simplified as;

$$v_{\mu v}^K \cong \sum_{A \in K} \langle \mu | (Q_A / | \mathbf{r} - \mathbf{r_A} |) | v \rangle \tag{22}$$

reducing the computational cost of the two-electron integrals by another factor of N_B.

Inter-fragment interactions have a similar approximation that evaluates the electrostatic contribution to the energy using the monomer densities for far separated dimers, instead of calculating the dimer density itself. This contribution is added to the dimer energy as

$$E'_{IJ} \cong E'_I + E'_J + Tr(D^I u^{1,I(J)}) + Tr(D^J u^{1,J(I)}) + \sum_{\mu\nu \in I} \sum_{\rho\sigma \in J} D^I_{\mu\nu} D^J_{\rho\sigma} (\mu\nu \mid \rho\sigma) \qquad (23)$$

where $u^{1,I(J)}$ and $u^{1,J(I)}$ are one-electron Coulomb potentials of the force exerted by fragment J on fragment I, and fragment I on fragment J, respectively.

Other approximations of the same nature are implemented for correlated dimers and trimers, where the corresponding corrections for far separated pairs and triples of fragments are neglected.[72,74]. Formal definitions and descriptions of the trimer interactions and cut-offs used in FMO3 have been described previously[69,70] and will not be discussed here. All of these approximations are based on a distance definition R_{app}, defined as the minimum distance between atoms in n-mer I and monomer J divided by the sum of their van der Waals radii.

There have been several new developments in the FMO theory that cannot be discussed in detail here. Nonetheless, it is useful to mention a few of them briefly. As an alternative to the original bond fragmentation scheme, in which the electron density describing the detached bonds is variationally optimized, a new scheme has been suggested in which this density is obtained for a model system and is kept frozen in fragment calculations[75]. This new scheme has been shown to work well for covalent crystals such as zeolites. The FMO method has also recently been implemented for the study of excited states[76], using multi-configurational self-consistent field (MCSCF)

theory, configuration interaction (CI), and time-dependent density functional theory (TDDFT).

4.2 FMO2 and FMO3 Calculations on $(H_2O)_{32}$ Clusters

The unusual characteristics of liquid water make it both very important to chemical processes and particularly difficult to model accurately. The structure of small $(H_2O)_n$ (n=6 through 20) clusters have recently been determined[83] using coupled cluster theory; however, the ability to model water clusters larger than this at the same level of theory is nearly impossible. The FMO method provides a way to model much larger water clusters at high levels of theory such as CCSD(T), while keeping the computational cost manageable.

In the present work, calculations of the energies of $(H_2O)_{32}$ water clusters are reported, using fully *ab initio* Møller-Plesset second order perturbation theory (MP2) as well as the MP2 implementation of the FMO method[72,77]. For these clusters a fragment (monomer) is defined as one water molecule for both FMO2 and FMO3 calculations. Initial structures were obtained from EFP Monte Carlo/Simulated Annealing (MC/SA) simulations followed by EFP optimizations of a representative set of structures. The MC/SA method with local minimization was used to sample the configuration space. For each global minimum found, the number of structures sampled was on the order of 500,000 – 1,100,000. The number of steps taken for each temperature was varied (100, 500, 1000, 10,000), along with changing the number of steps between local minimizations (10, 100, 1,000). The number of fragments moved per step was also varied between one and five. The starting temperature for the simulated annealing varied from 500 to 20,000 K and the final temperature was kept at 300 K. This selection of isomers

(the lowest energy structure is shown in Figure 7) was used to investigate the accuracy of the FMO method by comparing both absolute and relative energies.

Average errors for the FMO2-MP2 calculations (Table 6) using the 6-31++G(d,p) basis set are very consistent, around 12 kcal/mol. The FMO3-MP2 results illustrate the importance of three-body interactions in water clusters,[78,79] again with very consistent errors of ~2-3 kcal/mol (Table 6). Comparing results between basis sets in Table 6, when the basis set size is increased to 6-311++G(3df,2p), the FMO2 errors double to ~24-28 kcal/mol, while the FMO3 errors are cut in half to ~1 kcal/mol. This increase in errors with an increased basis set size for the two-body FMO method could be due to an increased importance of three-body contributions when the better basis set is used. The larger basis set also provides a better description of three-body interactions, making the lack of these interactions in FMO2 even more detrimental.

Despite the large absolute errors present in the FMO2 description of water clusters, the relative energetics of the different isomers is captured quite well. On average, the FMO2 relative energies are in agreement with full *ab initio* results to within ~1-2 kcal/mol with both basis sets, shown in Table 7. The error increases for FMO2 as the relative energy of the isomers increases, showing an increased importance of three body contributions with higher energy isomers. For both basis sets, the FMO3 results are within ~0.5 kcal/mol or less for all isomers as shown in Table 7, effectively removing the error from the two body description used in FMO2.

4.3 The FMO Method Applied to Ionic Liquids

Previous studies of ionic liquids[80,81,82] have focused on the decomposition of ion pairs (Figure 8), providing insight into the chemistry of their ignition as high energy

fuels. The focus of this paper, however, will be to accurately describe larger systems beyond single anion-cation pairs. Recent work by Li et. al.[83] has provided an accurate structure of two ion pairs (two cations and two anions), providing a greater understanding of the molecular structure and intermolecular interactions. The same system will be modeled here, along with systems of three ion pairs to illustrate the effectiveness of the FMO method in accurately describing complex molecular clusters, with the goal of modeling much larger systems in the future.

Two ionic liquid systems, 1-H,4-H-1,2,4-triazolium dinitramide and 1-amino, 4-H-1,2,4-triazolium dinitramide (Figure 8), were studied using both *ab initio* Møller-Plesset second order perturbation theory (MP2) and the MP2 implementation of the FMO method[72,77] with one ion chosen as an FMO fragment or monomer. Structures composed of two cations and two anions (tetramers), shown in Figures 9 and 11, and larger clusters of three cations and three anions (hexamers), shown in Figures 10 and 12, were optimized at the MP2/6-31+G(d) level of theory. FMO2-MP2 and FMO3-MP2 single point energy calculations were then performed for comparison with the fully *ab initio* results. Mulliken charges on each cation and anion were also compared to ensure that the pronounced charge separation present in ionic liquids[80,81,82] is captured using the FMO method.

Comparing the energies from FMO2 and FMO3, it can be seen immediately that the FMO method captures the total energy very well, within 2 kcal/mol in the worst case (Table 8). For the tetramers, both FMO2 and FMO3 are in good agreement; the FMO2 errors are less than 1 kcal/mol relative to the fully *ab initio* results. For the hexamers, the FMO2 errors are less than 2 kcal/mol, illustrating that FMO3 is not required to achieve the desired level of accuracy for these particular ionic liquid systems. Whether this trend

persists as system size grows beyond three ion pairs, or for other ion pairs, must be tested further.

As shown in previous studies[80,81,82], ionic liquid ion pairs have a definite separation of charge (approximately +1 on the cations, -1 on the anions) at equilibrium geometries. This charge separation is also observed for tetramers, as shown in Table 2, and the charge separation between cations and anions is still present up to hexamer structures. FMO2 captures the qualitative charge separation quite well, however, the magnitude of charge present on both cations and anions is slightly overestimated by FMO2 for both tetramer structures (Table 9). However as the system size increases to three ion pairs, the difference between FMO2, FMO3 and *ab initio* results becomes minimal. Future work using larger basis sets will help determine if FMO2 is accurate enough to describe larger ionic liquid clusters or if FMO3 will be required.

Another consideration for larger molecular systems is the computer time required. To illustrate the overall effectiveness of the FMO method in both providing accurate results and reducing computational requirements, timings were performed for the ionic liquid systems described above. Due to the fact that FMO2 is in good agreement with the *ab initio* results, only timings for FMO2 will be shown. However, it is noted here that because the tetramers and hexamers examined here are small, the FMO3 timings for these systems do no exhibit any time savings relative to the full MP2 calculations. The benefit of using FMO3 is only seen with larger systems[73].

Timings were performed on a Cray XT4 supercomputer using AMD Opeteron64 processors running at 2.1 GHz, located at the U.S. Army Engineer Research and Development Center (ERDC) in Vicksburg, Mississippi. Single point Møller-Plesset

second order perturbation theory (MP2) energy calculations were performed using 8, 16 and 32 processors with both FMO2 and MP2 using the 6-31+G(d) basis set. As shown in Table 10, FMO2 requires approximately half the computer time of a full MP2 calculation on the tetramers. With the increase in available processors, the overall time requirements are cut in half for both FMO2 and MP2, showing good scalability for both methods. With an increase in system size from ionic liquid tetramers to hexamers, the computer time required for a fully *ab initio* calculation increases more than 6 fold, while the FMO2 requirement only doubles. So, the FMO2 cost savings relative to full MP2 is much greater than that observed for the tetramers. Again, scalability for both methods is very good for the hexamers, cutting the computational time in half when doubling the number of available CPUs.

It is apparent that as the system size increases to larger ionic liquid clusters, or as the basis set size increases (or both) the computational requirements for a fully *ab initio* calculation will rapidly and increasingly exceed those for FMO2. It may be that as the system size increases, the importance of three-body contributions to the interaction energy will also increase, requiring the use of FMO3. Future work will determine the importance of three-body terms in ionic liquid systems, as well as the ability of the FMO method to describe larger molecular clusters.

5. Summary and Conclusions

Obtaining chemical accuracy (1 kcal/mol) using model chemistries has been a major focus of quantum chemistry research for the last quarter of a century. The desire to study larger systems in order to capture novel chemical phenomena (e.g. solvent effects, surface science, enzyme and heterogeneous catalysis and polymerization phenomena),

including the kinetics and dynamics of such processes, often requires very accurate predictions of potential energy surfaces for subsequent predictions to be even qualitatively correct. The computational effort of traditional methods such as Hartree-Fock (HF), density functional theory (DFT), 2nd order perturbation theory (MP2), and coupled cluster theory with perturbative triples (CCSD(T)) scale as $O(n^4)$, $O(n^4)$, $O(n^5)$, and $O(n^7)$, respectively, where n represents the size of the system, e.g., the size of the basis set. In practice, this limits the sizes of systems that can be studied with HF/DFT, MP2 and CCSD(T) to approximately a few hundred, one hundred, and twenty non-hydrogen atoms, respectively. By developing highly parallel algorithms, the goal of using sophisticated electronic structure methods to investigate large molecular problems becomes more feasible, especially if one has access to massively parallel computing hardware. However, scalability beyond hundreds to a few thousand processors is generally a serious bottleneck for correlated electronic structure methods. Consequently, parallel computing is not the sole solution to enabling accurate calculations on extended molecular systems; other approaches are needed. If one is interested in performing long-time simulations at reliable levels of theory, the situation is only exacerbated.

Pioneering work by Warshel[9a] and others[9] introduced hybrid methods that employ both quantum mechanics (QM) and molecular mechanics (MM), leading to the now ubiquitous QM/MM approach. Importantly, the QM/MM approach is quite general, so it can be employed with any level of QM, including the fragmentation methods that have been the primary focus of the present work. Modern fragmentation methods have their roots in ideas from Murrel (1955)[46a] and Christoffersen (1971).[46b] More recently developed fragmentation methods, such as the fragment molecular orbital (FMO) method

and the systematic fragmentation method (SFM), are now becoming capable of achieving chemical accuracy for extended molecular systems.

The effective fragment potential (EFP)[12] method has been developed to model non-bonded, intermolecular interactions. There are two related implementation of the EFP method: The original method, called EFP1, was developed specifically to study aqueous solvent effects on chemical processes. The more recently developed EFP2 method is completely general, in the sense that an EFP2 contains no empirically fitted parameters. The Coulomb and induction terms are common to EFP1 and EFP2, and the remaining terms in EFP2 are derived from first principles. Once an EFP2 has been built for a specific system (accomplished by a straightforward GAMESS run), the evaluation of EFP-EFP interactions requires a small fraction of the computational cost compared to the fully QM calculation. The EFP computational effort scales in the range of quadratic to linear with an increasing number of fragments. EFP1/MP2 achieves an accuracy of ~1 kcal/mol for the relative energies of six-water clusters compared to CCSD(T)/aug-cc-pVTZ.[47b] For benzene dimer binding energies, EFP2 achieves an accuracy of ~1 kcal/mol relative to CCSD(T)/aug-cc-pVTZ results. The EFP1 method has been interfaced with the QM methods HF,[13] DFT,[14] MCSCF,[15] singly excited configuration interaction (CIS),[16] and time-dependent density functional theory (TDDFT)[17] within the GAMESS suite, so EFP1 is a fully QM/MM method. The EFP2-QM integration is currently under development.[18] These new features will greatly expand the utility of the method, by enabling, for example, the exploration of solvent effects for a wide variety of problems in organic and inorganic chemistry.

The SFM fragments a molecule based on the number of single bonds in each

fragment, while considering the environmental effects of distant parts of the system via a many body expansion of the interactions not captured by the internal energies of the fragments. This framework allows the SFM to be widely applicable with a simple user interface, which has been integrated into the GAMESS suite. The SFM has been used to study small and medium sized organic molecules[35], as well as crystals.[37] In this paper, it was demonstrated that the SFM, when using EFPs for the non-bonded interactions, has a mean averaged error of 1.8 kcal/mol for several α-helical isomers at the HF/6-31++G(d,p) level of theory. The SFM is formally independent of the *ab initio* methods used in calculations of the fragments, thereby facilitating highly accurate energies and relative energies with nearly linear scaling as the size of the system is increased. Therefore, the SFM can be used in concert with any available electronic structure method, such as MP2 and CCSD(T), and applied to much larger molecular systems that might otherwise be accessible. The time requirements for the EFP part of a SFM calculation, when EFPs are used for the non-bonded interactions, are determined by the cost of an initial HF single point calculation that is employed to generate the potential. Therefore, the EFP fraction of the overall computer time requirements decreases rapidly as the level of *ab initio* theory increases (e.g., from HF to MP2 to CCSD(T)).

The FMO method treats each fragment (monomer, dimer, etc.) in a Coulomb bath that represents the remainder of the full system. The energy of each monomer is iterated to self-consistency within this Coulomb bath. The FMO method is very flexible with regard to the definition of fragments (i.e., monomers), the assignments of distance cutoff parameters, and the level of many-body effects (i.e., dimer, trimer, etc.) to be included in the calculation. Combined with the avoidance of capping procedures, this facilitates the

study of a wide variety of systems including clusters, zeolites, and proteins, and the ability to balance accuracy and computational efficiency. Within GAMESS, the FMO method has been interfaced with the polarizable continuum method and the EFP method for studies of solvent effects on chemical processes. Each monomer in a molecular system of interest can be treated by most traditional electronic structure methods. In the present work, the FMO method has been shown to achieve accuracy within 1 kcal/mol for both ionic liquid systems and water clusters.

The EFP method provides a systematically improvable description of non-bonded interactions, while the FMO method and the SFM facilitate the description of large molecular systems with high levels of accuracy. The interface of the two fragmentation methods for internal and near-field *ab initio* calculations with the EFP method for non-bonded moderate and far-field interactions and for solvent effects provides a powerful and computationally effective combination. Additionally, the ability of these methods to take advantage of the standard theoretical electronic structure framework allows their capabilities to move forward with new advances in electron correlation, wavefunction description and basis set development for large molecular systems. The primary limitation of both the SFM and FMO methods is that they are primarily applicable to "localized" systems. That is, these methods rely on the ability to decompose a large species into smaller fragments that are reasonably self-contained. So, the methods would not work well for highly delocalized systems, such as a conducting metal, graphite, or a linear polyene.

The SFM and FMO methods have not yet been broadly applied to the study of chemical reactions. Since analytic gradients are available for both methods, the

exploration of potential energy surfaces for chemical reaction of complex systems using these methods is a logical next step.

An important advantage of the FMO and SFM approaches described here is their ability to take great advantage of massively parallel computers. Because the energy of each fragment can be computed essentially independently, each fragment calculation can be determined on a separate compute node. Further, because most of the algorithms used in GAMESS for electronic structure functionalities are themselves highly scalable, the fragment-based calculations can take advantage of multi-level parallelism. This capability, which is enhanced by middleware developments like the generalized distributed data interface (GDDI), bodes well for the implementation of algorithms for "petascale" computers that are expected to come on line within the next 2-3 years. Simultaneous advancements in new approaches like the fragmentation methods discussed here, novel parallel algorithms, *ab initio* theory, and novel approaches in hardware development are all required if one is to successfully address the grand challenge problems in the chemical sciences, biological sciences and materials science and engineering.

References

1. (a) G4: Curtiss L.A.; Redfern, P.C.; Raghavachari, K. *J.Chem. Phys.* **2007**,*127*, 124105. (b) G4: Curtiss L.A.; Redfern, P.C.; Raghavachari, K. *J.Chem. Phys.* **2007**, *126*, 084108. (c) G3SX: Curtiss L.A.; Raghavachari, K; Redfern, P.C.; Pople. J. A. *J.Chem. Phys.* **2001,** *114*, 108. (d) G3: Curtiss L.A.; Raghavachari, K; Redfern, P.C. *J.Chem. Phys.* **1998**, *109*,7764. (e) G3(MP2): Curtiss L.A.; Redfern, P.C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703. (f) G2(MP2): Curtiss, L. A.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys*. **1993**, *98*, 1293. (g) G2: Curtiss L.A.; Raghavachari, K; Trucks, G. W.; Pople. J. A. *J.Chem. Phy.* **1991,** *94*,7221. (h) G1: Curtiss L.A.; Jones, C.; Trucks, G. W.; Raghavachari, K; Pople. J. A. *J.Chem. Phys.* **1990,** *93*, 2537. (i) G1: Pople. J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K; Curtiss L.A. *J.Chem. Phys.* **1989.** *90*, 5622.

2. (a) W4: Karton, A.; Rabinovich, E.; Martin, J.M. L.; Ruscic, B. *J.Chem. Phys.* **2006**, *125*, 144108. (b) W3: Boese, A. D.; Oren, M.; Atasoylu, O.; Martin, J.M. L.; Kállay, M.; Gauss, J. *J.Chem. Phys.* **2004,** *120*, 4129. (c)W1&W2: Martin, J.M. L.; Oliveira, G**.** *J.Chem. Phys.* **1999,** *111*, 1843.

3. (a) ccCA: DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *124*, 114103. (b) ccCA: DeYonker, N. J.; Grimes, T.; Yockel, S.; Dinescu, A.; Mintz, B.; Cundari, T. R.; Wilson, A. K. *J. Chem. Phys.* **2006**, *125*, 104011. (c) ROCBS-QB3: Wood, G. P. F.; Random, L.; Petersson, G. A.; Barnes, E. C.; Frisch, M. J.; Montgomery, J. A. *J. Chem. Phys.* **2006**, *125*, 094106. (d) CBS-4M: Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **2000**, *112*, 6532. (e) CBS-QB3: Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. *J. Chem. Phys.* **1999**, *110*, 2822. (f) SAC: Gordon, M.; Truhlar, D. G. *J. Am. Chem. Soc.* **1986**, *108*, 5412. (g) SAC: Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *234*, 64. (h) SAC: Fast, P. L.; Corchado, J. C.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 5129. (i) SEC: Corchado, J. C.; Truhlar, D. G. *ACS Symp. Ser.* **1998***, 712*, 106. (j) SEC: Brown, F. B.; Truhlar, D. G. *Chem. Phys. Lett.* **1985**, *117*, 307. (k) HEAT: Szalay, P. G.; Tajti, A.; Stanton, J. F. *Mol. Phys.* **2005**, *103*, 2159. (l) HEAT: Tajti, A.; Szalay, P. G.; Csa´ sza´ r, A. G.; Kallay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vazquez, J.; Stanton, J. F. *J. Chem. Phys.* **2004**, *121*, 11599. (m) Kedziora, G. S.; Pople, J. A.; Rassolov, V. A.; Ratner, M. A.; Redfern, P. C.; Curtiss, L. A. *J. Chem. Phys.* **1999**, *110*, 7123. (n) Cioslowski, J.; Schimeczek, M.; Liu, G; Stoyanov, V. *J. Chem. Phys.* **2000**, *113*, 9377. (o) Grimme, S. *J. Chem. Acc.* **1992**, *81*, 405. (p) Ricca, A.; Bauschlicher, C. W. *J. Phys. Chem. A* **1998**, *102*, 876. (q) Chase, J. M. W. NIST-JANAF Tables, 4th ed. *J. Phys. Chem. Ref. Data Monogr.* **1998**, *9*, 1.

4. (a) Almlöf, J.; Taylor, P.R. *J. Chem. Phys*. **1987**, *86*, 4070. (b)Almlöf, J.; Helgaker, T.; Taylor, P.R.; *ibid*, **1988**, *92*, 3029.

5. (a) Dunning ,T. H; Jr., *J. Chem. Phys*. **1989**, *90*, 1007. (b) S. S. Xantheas; T. H. Dunning, Jr. *J. Phys. Chem*. **1993**, *97*, 18. (c) T. H. Dunning, Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244. (d) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; T. H. Dunning, Jr. *J. Chem. Phys.* **1999** , *110*, 7667. (e)

Wilson, A. K.; Woon, D. E.; Peterson, K. A.; T. H. Dunning, Jr. *J. Abstr. Pap.-Am. Chem. Soc*. **1997**, *213*, 60. (f) Kendall, R. A.; T. H. Dunning, Jr.; Harrison, R. J. *J. Chem. Phys*. **1992**, *96*, 6796. (g) Woon, D. E.; T. H. Dunning, Jr. *J. Chem. Phys*. **1994**, *100*, 2975. (h) Woon, D. E.; T. H. Dunning, Jr. *J. Chem. Phys*. **1995**, *103*, 4572. (i) Peterson, K. A.; T. H. Dunning, Jr. *J. Chem. Phys*. **2002**, *117*, 10548. (j) G3 Basis: Curtiss, L. A.; Redfern, P. C.; Rassolov, V.; Kedziora, G.; Pople, J. A. *J. Chem. Phys*. **2001**, *114*, 9287.

6. Peterson, K.A.; Adler, T.B.; Werner, H.-J. *J. Chem. Phys*. **2008**, *128*, 084102.

7. (a) Silagen: Redondo, A.; Goddard, W. A. III. *J. Vac. Sci. Technol*. **1982**, *21*, 344. (b) SM8: Cramer, C. J.; Truhlar, D. G. *Acct. Chem. Res*. **2008**, *41*, 760. (c) COSMO: Klamt, A.; Schuurmann, G. *J. Chem. Soc., Perkin Trans*. **1993**, *11*, 799. (d) PCM: Tomasi, J.; Persico, M. *Chem. Rev*. **1994**, *94*, 2027. (e) GCOSMO: Truong, T.N.; Stefanovich, E.V. *Chem. Phys. Lett*. **1995**, *240*, 253.

8. (a) Centroid MD: Jang, S.; Voth, G. A. *J. Chem. Phys*. **1999**, *111*, 2371. (b) Force Matching: Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys*. **2004**, *120*, 10896. (c) Izvekov, S.; Voth, G. A. *J. Chem. Phys*. **2005**, *123,* 044505. (d) Kuwajima S.; Warshel, A. *J. Chem. Phys*. **1988**, *89*, 3751. (e) Warshel; A.; Hwang, J-K. *J. Chem. Phys*. **1986**, *84*, 4938.

9. (a) Warshel, A.; Levitt M.; *J. Mol. Biol*. **1976**,*103*, 227. (b) Gao, J.; Truhlar, D. G. *Annu. ReV. Phys. Chem*. **2002**, *53*, 467. (c) Friesner, R. A.; Guallar, Y. *Annu. ReV. Phys. Chem.* **2005**, *56*, 389. (d) Damjanovic, A.; Kosztin, I.; Kleinekathoefer, U.; Schulten, K. *Phys.Revi. E,* **2002**, *65*, 031919. (e) Chen, J.; Martínez, T.J. *Chem. Phys. Lett*. **2007**, *438*, 315. (f) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem*. **1990**, *11*, 700. (g) Stanton, R. V.; Little, L. R.; Merz, K. M. *J. Phys. Chem*. **1995**, *99*, 17344. (h) Froese, R. D. J.; Morokuma, K. *Chem. Phys. Lett*. **263**, *393,* 1996. (i) Gao, J. L. *Acc. Chem. Res*. **1996**, *29*, 298. (j) Singh, U. C.; Kollman, P. A. *J Comput Chem* **1986**, *7*, 718. (k) Gao, J. In Reviews in Computational Chemistry, Vol. 7; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, **1996**, p. 119. (l) Singh, U. C.; Kollman, P. A. *J Comput Chem* **1986**, *7*, 718. (m) Mordasini, T. Z.; Thiel, W. Chimia **1998**, *52*, 288.

10. (a) Svensson, M.; Humbel, S.; Froese, R. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357. (b) Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, *24*, 760. (c) Canfield, P.; Dahlbom, M. G.; Hush, N. S.; Reimers, J. R. *J. Chem. Phys.* **2006**, *124*, 024301.

11. (a) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 3898. (b) Fast, P. L.; Corchado, J. C.; Sanchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 5129. (c) Fast, P. L.; Sanchez, M. L.; Corchado, J. C.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *110*, 11679. (d) Fast, P. L.; Sanchez, M. L.; Truhlar, D. G. *Chem. Phys. Lett*. **1999**, *306*, 407. (e) Tratz, C. M.; Fast, P. L.; Truhlar, D. G. *PhysChemComm* **1999**, *2*, 14. (f) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys*. **2000**, *112*, 1125. (g) Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 6111. (h) Curtiss, L. A.; Redfern, P. C.; Rassolov, V.; Kedziora, G.; Pople, J. A. *J. Chem. Phys*. **2001**, *114*, 9287. (i) Lee, T.-H.; Chen, H.-R.; Hu, W.-P. *Chem. Phys. Lett.* **2005**, *412*, 430. (j) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43. (k) Sun, Y.-L.; Lee, T.-H.;

Chen, J.-L.; Wu, K.-J.; Hu, W.-P. *Chem. Phys. Lett.* **2007**, *442*, 220. (l) Chen, J.L.; Sun, Y.L.; Wu,K.J; Hu, W.P. *J. Phys. Chem. A* **2008**, *112*, 1064.

12. (a) Gordon, M.S.; Slipchenko, L.; Hui, L.; Jensen, J.H. *Ann. Rep. Comp. Chem.* **2007**, *3*, 177. (b) Gordon, M.S.; Freitag, M.A.; Bandyopadhyay, P.; Kairys, V.; Jensen, J.H.; Stevens, W.J. *J. Phys. Chem.* **2001**, *105*, 293.

13. Day, P.N.; Jensen, J.H.; Gordon, M.S.; Webb, S.P.; Stevens, W.J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.

14. Adamovic, I.; Gordon, M.S. *J. Phys. Chem.* **2005**, *109*, 1629.

15. Webb, S.P.; Gordon, M.S. *J. Phys. Chem. A* **1999**, *103*, 1265.

16. Pooja, A.; Gordon, M. S.; in preparation

17. Yoo, S.; Zahariev, F.; Sok, S.; Gordon, M. S.; *J. Chem. Phys.* **2008**, *129*, 14112.

18. (a) Kemp, D. D.; Gordon, M. S. manuscript in preparation. (b) Smith, T.; Slipchenko, L.; Ruedenberg, K; Gordon, M. S. manuscript in preparation.

19. (a) Stone, J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047. (b) Stone, A.J., The Theory of Intermolecular Forces; Oxford Press; 1996. (c) Stone, A. J. *J. Chem. Theory Comp.* **2005**, *1*, 1128.

20. (a) Edmiston C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457. (b) Raffenetti, R. C.; Ruedenberg, K.; Janssen, C. L.; Schaefer, H. F. III. *Theor. Chim. Acta* **1993**, *86*, 149.

21. (a) Slipchenko, L.; Gordon, M.S. *J. Comp. Chem.* **2006**, *28*, 276. (b) Freitag, M.A.; Gordon, M.S.; Jensen, J.H.; Stevens W.J. *J. Chem. Phys.* **2000**, *112*, 7300.

22. (a) Slipchenko, L.; Gordon, M. S.; in press. (b) Murrell, J. N.; Teixeira-Dias, J. J. C. *Mol. Phys.* **1970**, *19*, 521.

23. (a) Jensen, J.H; Gordon, M.S. *Mol. Phys.* **1996**, *89*, 1313. (b) Murrell, J. N. *Proc. Roy. Soc. (Lond.)* **1965**, *A284*, 566. (c) Murrell J. N. *J. chem. Phys.* **1967**, *47*, 4916.

24. (a) Adamovic, I.; Gordon, M.S. *Mol. Phys.* **2005**, *103*, 379. (b) Amos, R.D.; Handy, N.C.; Knowles, P.J.; Rice, J.E.; Stone, A.J. *J. Phys. Chem.* **1985**, *89*, 2186. (c) Piecuch, P. *Mol. Phys.* **1986,** 59, 1085.

25. Tang, K.T.; Toennies, J.P. *J. Chem. Phys.* **1984**, *80*, 3726.

26. Li, H.; Gordon, M.S.; Jensen, J.H. *J. Chem. Phys.* **2006**, *124*, 214107.

27. Li, H.; Gordon, M.S., *Theor. Chem. Accts.* **2006**, *115*, 385.

28. (a) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599. (b) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039. (c) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. *Chem. Phys. Lett.* **2004**, *394*, 293. (d) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2004,** *120*, 839. (e) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145. (f) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 11386 (g) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903. (h) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Comput. Chem.* **2004**, *25*, 1431. (i) Chen, X.; Zhang, Y.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105. (j) Zhang, D. W.; Zhang, J. Z. H. Int. *J. Quantum Chem.* **2005**, *103*, 246. (k) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2. (l) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7251. (m) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112. (n) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103. (o) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *124*, 184703. (p) Mei,

Y.; Ji, C.; Zhang, J. Z. H. *J. Chem. Phys*. **2006**, *125*, 094906. (q) Mei, Y.; Wu, E. L.; Han, K. L.; Zhang, J. Z. H. *Int. J. Quantum Chem*. **2006**, *106*, 1267. (r) Mei, Y.; He, X.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins- Struct. Funct. Bioinform*. **2005**, *59*, 489. (s) He, X.; Mei, Y.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins- Struct. Funct. Bioinform*. **2005**, *61*, 423.

29. (a) Imamura, A.; Aoki, Y.; Maekawa, K. *J. Chem. Phys*. **1991**, *95*, 5419. (b) Aoki, Y.; Suhai, S.; Imamura, A. *Int. J. Quantum Chem*. **1994**, *52*, 267. (c) Aoki, Y.; Suhai, S.; Imamura, A. *J. Chem. Phys*. **1994**, *101*, 10808. (d) Mitani, M.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem*. **1995**, *54*, 167; (e) Mitani, M.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem*. **1997**, *64*, 301. (f) Rather, G.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem.* **1999**, *74*, 35. (g) Gu, F. L.; Aoki, Y.; Imamura, A.; Bishop, D. M.; Kirtman, B. *Mol. Phys*. **2003**, *101*, 1487. (h) Gu, F. L.; Aoki, Y.; Korchowiec, J.; Imamura, A.; Kirtman, B. *J. Chem. Phys*. **2004**, *121*, 10385. (i) Korchowiec, J.; Gu, F. L.; Imamura, A.; Kirtman, B.; Aoki, Y. *Int. J. Quantum Chem*. **2005**, *102*, 785. (j) Korchowiec, J.; Gu, F. L.; Aoki, Y. *Int. J. Quantum Chem*. **2005**, *105*, 875. (k) Makowski, M.; Korchowiec, J.; Gu, F. L.; Aoki, Y. *J. Comput. Chem*. **2006**, *27*, 1603. (l) Orimoto, Y.; Aoki, Y. J. *Polym. Sci. B: Polym. Phys*. **2006**, *44*, 119.

30. (a) Gadre, S. R; Shirsat, R. N.; Limaye, A. C.; *J. Phys. Chem*. **1994**, *98*, 9165. (b) Babu, K.; Gadre, S. R.; *J. Comput. Chem*. **2003**, *24*, 484. (c) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E.; *Theor. Chem. Acc*. **2004**, *111*, 255. (d) Gadre, S. R.; Ganesh, V.; *J. Theor. Comp. Chem*. *2006, 5, 835.*

31. (a) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys*. **2005**, *103*, 2255. (b) Kamiya, M.; Hirata, S.; Valiev, M. *J. Chem. Phys*. **2008**, *128*, 074103. (c) Windus ,T.L.; Bylaska, E.J.; Dupuis, M.; Hirata, S.; Pollack, L.; Smith, D.M.; Straatsma, T.P.; Aprà, E. In Lecture Notes in Computer Science, vol. 2660, P.M.A. Sloot, D. Abramson, A. Bogdanov, J.J. Dongarra, A. Zomaya, Y. Gorbachev (Eds), Springer, Berlin (2003). (d) Hirata. S. *J. phys. Chem. A* **2003**, *107*, 9887. (e) Hirata, S. *J. Chem. Phys*. **2004**, *121*, 51. (f) Hirata, S.; Zhan, C.-G.; Aprà, E.; Windus, T.L.; Dixon, D.A. *J. Phys. Chem. A* **2003**, *107*, 10154.

32. (a) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput*. **2007**, *3*, 46. (b) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput*. **20078**, *4*, 1.

33. (a) Tschumper, G. S.; *Chem. Phys. Lett*. **2006**, *427*, 185. (b) Hopkins, B. W.; Tschumper, G. S. *Mol. Phys*. **2005**, *103*, 309. (c) Hopkins, B. W.; Tschumper, G. S. *Chem. Phys. Lett.* **2005**, *407*, 362.

34. Deev, V.; Collins, M.A. *J. Chem. Phys*. **2005**, *122*, 154102.

35. Collins, M A; Deev V A. *J. Chem. Phys*. **2006**, *125*, 104104.

36. Collins, M.A. *J. Chem. Phys*. **2007**, *127*, 024104.

37. Netzloff, HM; Collins, M.A. *J. Chem. Phys*. **2007**, *127*, 134113.

38. Mullin, J.M.; Collins,M.A.; Gordon,M.S. In preparation.

39. Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett*. **1999**, *312*, 319.

40. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett*. **1999**, *313*, 701.

41. Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2000**, *318*, 614

42. Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett.* **2001**, *336*, 163.

43. Inadomi, Y.; Nakano, T.; Kitaura, K.; Nagashima, U. *Chem. Phys. Lett.* **2002**, *364*, 139.

44. Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2002**, *351*, 475.

45. Fedorov, D. G.; Kitaura, K.. *J.Phys. Chem. A* **2007**, *111*, 6904.

46. (a) Christoffersen, R.; Genson,D. W.; Maggiora, G. M.; *J. Chem. Phys.* **1971**, *54*, 239. (b) Longuet-Higgins, H. C.; Murrel, J. N.; *Proc Phys. Soc*, **1955**, *A68*, 601.

47. (a) Curtis, L;. Janssen, I.; Nielsen, M. B. *Parallel Computing in Quantum Chemistry*, CRC press, 2008. (b) Olson, R.M.; Bentz, J. L.;. Kendall, R. A.; Schmidt, M. W.; Gordon, M.S. *J. Chem. Theory Comput.* **2007**, *3*, 1312. (c) Bentz, J. L.; Olson, R. M.; Gordon, M. S.; Schmidt, M. W.; Kendall, R. A. *Comp. Phys. Com.* **2007**, *176*, 589. (d) Dudley, T.J.; Olson, R.M.; Schmidt, M.W.;Gordon, M.S. *J. Comp. Chem.* **2006**, *27*, 353. (e) Gordon, M.S.; Ruedenberg, K.; Schmidt, M.W.; Bytautas, L.; Dudley, T.J.; Nagata, T.; Olson, R.; Varganov, S. *J. Phys. Conf. Ser.* **2006,** *46*, 229. (f) Janowski, T.; Ford, A. R.; Pulay, P. *J. Chem. Theory Comput.* **2007**, *3*, 1368.

48. (a) Fletcher, G.D.; Schmidt, M.W.; Gordon, M.S. *Adv. Chem. Phys.* **1999**,*110*, 267. (b) Schmidt, M.W.; Fletcher, G.D.; Bode, B.M.; Gordon, M.S. *Comp. Phys.Comm.* **2000**,*128*, 190. (c) Olson, R.M.; Schmidt, M.W.; Gordon, M.S.; Rendell, A.P.; *Proc. Supercomputing* **2004**. (d) Nieplocha, J.; Harrison, R. J.; Littlefield, R. Proc. Supercomputing 1994; IEEE Computer Society Press: Washington, D.C., **1994**, pp 340−346. (e) Nieplocha, J.; Palmer, B.; Tipparaju, V.; Manojkumar, K.; Trease, H.; Aprà, E. *Int. J. High Perform, Comput. Appl.* **2006**, *20*, 203.

49. Fedorov, D.G.; Olson, R.M.; Kitaura, K.; Gordon, M.S.; Koseki, S. *J. Comp. Chem.* **2004**, *25*, 872.

50. (a) Kemp, D.D. Gordon, M.S. *J. Phys. Chem A* **2005**, *109*, 7688. (b) Aikens, C.M.; Gordon, M.S. *J. Am. Chem. Soc.,* **2006**, 128, 12835. (c) Adamovic, I.; Gordon, M.S. *J. Phys. Chem, A* **2006**, *110*, 10267. (d) Adamovic, I.; Gordon, M.S. *J. Phys. Chem.* **2005**, *109*, 1629. (e) Freitag, M.A.; Hillman, B.; Gordon, M.S. *J. Chem. Phys.* **2004**, *120*, 1197. (f) Bandyopadhyay, P.; Mennucci, B.; Tomasi, J.; Gordon, M.S. *J. Chem. Phys.* **2002**, *116*, 5023. (g) Webb, S.P.; Gordon, M.S. *J. Phys. Chem. A* **1999**, *103*, 1265. (h) Mullin, J; Gordon, M.S. (in preperation)

51. (a) Netzloff, H.M.; Gordon, M.S. *J. Chem. Phys.* **2004**,*121*, 2711. (b) Li, H.; Netzloff, H.M.; Gordon, M.S. *J. Chem. Phys.* **2006**, *125*, 194103.

52. (a) Hehre, W.J.; Ditchfield, R.; Pople, J.A. *J. Chem. Phys.* **1972**, *56*, 2257 (b) Francl, M.M.; Pietro, W.J.; Hehre, W.J.; Binkley, J.S.; Gordon, M.S.; DeFrees, D.J.; Pople, J.A. *J.Chem.Phys*. **1982**, *77*, 3654. (c) Hariharan, P.C.; Pople, J.A. *Theoret. Chim. Acta* **1973**, *28*, 213.

53. (a) Hariharan, P. C.; Pople, J. A. *Theoret. Chim. Acta* **1973**, *28*, 213. (b) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650. (c)

Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. *J. Comp. Chem.* **1983**, *4*, 294.

54. (a) 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.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem.* **1993**, *14*, 1347. (b) Gordon, M.S.; Schmidt, M.W., Theory and Applications of Computational Chemistry, the first forty years; Elsevier; Amsterdam; 2005

55. (a) Hobza, P., Selzle, H.L., Schlag, E.W. *J. Am. Chem. Soc.* **1994**, *116*, 3500. (b) Hobza, P., Selzle, H.L., Schlag, E.W. *J. Phys. Chem.* **1996**, *100*, 18790. (c) Jaffe, R.L., Smith, G.D. *J. Chem. Phys.* **1996**, *105*, 2780. (d) Park, Y.C., Lee, J.S. *J. Phys. Chem. A* **2006**, *110*, 5091. (e) Spirko, V., Engvist, O., Soldan, P., Selzle, H.L., Schlag, E.W., Hobza, P. *J. Chem. Phys.* **1999**, *111*, 572. (f) Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M., Tanabe, K. *J. Am. Chem. Soc.* **2002**, *124*, 10. (g) Tsuzuki, S., Uchimaru, T., Sugawara, K., Mikami, M. *J. Chem. Phys.* **2002**, *117*, 11216. (h) Bornsen, K.O., Selzle, H.L., Schlag, E.W. *J. Chem. Phys.* **1986**, *85*, 1726. Felker, P.M., Maxton, P.M., Schaeffer, M.W. *Chem. Rev.* **1994**, *94*, 1787. (i) Janda, K.C., Hemminger, J.C., Winn, J.S., Novick, S.E., Harris, S.J., Klemperer, W. *J. Chem. Phys.* **1975**, *63*, 1419. (j) Law, K.S., Schauer, M., Bernstein, E.R. *J. Chem. Phys.* **1984**, *81*, 4871. (k) Scherzer, W., Kratzschmar, O., Selzle, H.L., Schlag, E.W. *Zeitschrift Fur Naturforschung Section Journal of Physical Sciences* **1992**, *47*, 1248. (l) Steed, J.M., Dixon, T.A., Klemperer, W. *J. Chem. Phys.* **1979**, *70*, 4940. (m) Puzder, A., Dion, M., Langreth, D.C. *J. Chem. Phys.* **2006**, *124,* 1674105. (n) Sato, T., Tsuneda, T., Hirao, K. *J. of Chem. Phys.* **2005**, *123*, 104307.

56. Sinnokrot, M.O.; Sherrill, C.D. *J. Phys. Chem. A* **2004**, *108*, 10200.

57. Sinnokrot, M.O.; Valeev, E.F.; Sherrill, C.D. *J. Am. Chem. Soc.* **2002**, *124*, 10887.

58. Arunan, E.; Gutowsky, H.S. *J. Chem. Phys.* **1993**, *98*, 4294.

59. Grover, J.R.; Walters, E.A.; Hui, E.T. *J. Phys. Chem.* **1987**, *91*, 3233.

60. Krause, H.; Ernstberger, B.; Neusser, H.J. *Chem. Phys. Lett.* **1991**, *184*, 411.

61. Jung, Y.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2831.

62. Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345.

63. Paldus, J., Wilson, S., editor. Handbook of Molecular Physics and Quantum Chemistry, vol. 2. Chichester: Wiley; 2000, p. 272–313.

64. Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.

65. Allen, F.H. *Acta Cryst.* **2002**, *B58*, 380.

66. Fedorov, D. G.; Kitaura, K.; Li, H.; Jensen, J. H.; Gordon, M. S. *J. Comp. Chem.* **2006**, *27*, 976.

67. Fedorov, D.G.; Ishida, T.; Kitaura, K. *J. Phys. Chem. A.* **2005**, *109*, 2638.

68. Mochizuki, Y.; Yamashita, K.; Murase, T.; Nakano, T.; Fukuzawa, K.; Takematsu, K.; Watanabe, H.; Tanaka, S. *Chem. Phys. Lett.* **2008**, *457*, 396.

69. Fedorov, D. G., Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832.

70. Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129.

71. a) Mulliken, R.S. *J. Chim. Phys.* **1949**, *46*, 500. b) Mulliken, R.S. *J. Chim. Phys.* **1949**, *46*, 521.

72. Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.

73. Yasuda, Y.; Yamaki, K. *J. Chem. Phys.* **2006**, *125*, 154101.

74. Fedorov, D.G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.

75. D. G. Fedorov, D.G.; Jensen, J.H.; Deka, R.C.; Kitaura, K. *J. Phys. Chem. A* **2008**, *112*, 11808.

76. (a) Fedorov, D.G.; Kitaura, K *J. Chem. Phys.* **2005**, *122*, 054108; (b) Mochizuki, Y; Koikegami, S; Amari, S; Segawa, K; Kitaura, K.; Nakano, T. *Chem. Phys. Lett.*, **2005**, *406,* 283; (c) Mochizuki, Y; Tanaka, K; Yamashita, K; Ishikawa, T; Nakano, T; Amari, S; Segawa, K; Murase, K; Tokiwa, H; Sakurai, M. *Theor. Chem. Acc.*, **2007**, *117*, 541; (d) Chiba, M; Fedorov, D.G.; Kitaura, K. *Chem. Phys. Lett*. **2007**, *444*, 346; (e) Chiba, M.; Fedorov, D.G.; Kitaura, K. *J. Chem. Phys.* **2007**, *127*, 104108; (f) Chiba, M.; Fedorov, D.G.; Kitaura, K. *J. Comp. Chem.* **2008**, *29*, 2667.

77. Fedorov, D. G.; Kitaura, K. *J. Comp. Chem*. **2007**, *28*, 222.

78. Fedorov, D.G.; Kitaura, K. *Chem. Phys. Lett*. **2006**, *433*, 182.

79. Xantheas, S. *Struct. Bond.* **2005**, *116*, 119

80. Schmidt, M.W., Gordon, M.S., Boatz, J.A. *J. Phys. Chem. A* **2005**, *109,* 7285.

81. Zorn, D.D., Boatz, J.A., Gordon, M.S. *J. Phys. Chem*. **2006**, *B110,* 11110.

82. Pimienta, I.S.O., Elzey, S., Boatz, J.A., Gordon, M.S. *J. Phys. Chem. A* **2007**, *111*, 691.

83. Li, H., Boatz, J.A., Gordon, M.S. *J. Am. Chem. Soc.* in press.

84. T. Nagata and M.S. Gordon, manuscript in preparation.

Table 1. Binding energies (kcal/mol) and equilibrium separations R (Å) of benzene dimer structures.

Method	Basis set	Sandwich		T-Shaped		Parallel-displaced		
		R	Energy	R	Energy	R_1	R_2	Energy
MP2[a]	aug-cc-pVDZ[b]	3.8	-2.83	5.0	-3.00	3.4	1.6	-4.12
	aug-cc-pVTZ	3.7	-3.25	4.9	-3.44	3.4	1.6	-4.65
	aug-cc-pVQZ[b]	3.7	-3.35	4.9	-3.48	3.4	1.6	-4.73
CCSD(T)[a]	aug-cc-pVDZ[b]	4.0	-1.33	5.1	-2.24	3.6	1.8	-2.22
	aug-cc-pVQZ[b]	3.9	-1.70	5.0	-2.61	3.6	1.6	-2.63
EFP	6-311++G(3df,2p)	4.0	-2.11	5.2	-2.50	3.8	1.2	-2.34

a Reference 56
b Basis sets as described in Ref. 56

Table 2. Mean absolute errors of isomerization energies (kcal/mol) calculated by SFM, relative to fully *ab initio* (no SFM) energies. The non-bonded terms use the combined EFP/*ab initio* approximation (cutoff ≥2.7Å). Given in parentheses is SFM with the non-bonded term fully *ab initio* (no approximation).

Isomer	2nd Order Many Body 6-31G(d,p)		3rd Order Many Body 6-31G(d,p)	
	HF kcal/mol	MP2 kcal/mol	HF kcal/mol	MP2 kcal/mol
ODETAS-AHALUQ	0.0 (0.5)	0.6 (0.1)	0.6 (0.2)	0.4 (0.3)
ODETAS01-AHALUQ	0.5 (1.2)	0.0 (0.7)	0.9 (0.4)	0.0 (0.1)
BAZGEP-BAZGIT	0.4 (0.6)	0.4 (0.2)	0.2 (0.5)	0.3 (0.3)
BELDIF-NOTGAE	2.6 (2.6)	4.6 (5.1)	2.6 (2.4)	4.6 (4.9)
FDOURD01-BOFWIC	0.5 (0.7)	0.2 (0.9)	0.1 (0.2)	0.5 (1.6)
CONBAI-FDMUPD10	0.7 (0.3)	1.1 (2.4)	3.5 (4.2)	2.9 (5.4)
IDUFES-IDUFAO	1.6 (2.1)	3.1 (5.3)	0.7 (0.2)	1.8 (1.1)
LEDRAN-LEDRER	1.0 (0.8)	1.0 (2.6)	1.7 (1.9)	1.6 (2.8)
LEDRIV-LEDRER	1.9 (1.5)	0.3 (0.0)	0.5 (0.4)	1.4 (1.0)
TAXYIA-MOGQOO	1.3 (0.4)	11.1 (8.4)	0.0 (1.2)	9.7 (6.8)
TAXYOG-MOGQOO	1.6 (2.1)	1.1 (3.7)	2.0 (2.6)	1.5 (4.2)
WINXIA-XEXXIH	0.3 (0.3)	0.1 (0.1)	0.3 (0.3)	0.1 (0.1)
MAE	1.0 (1.1)	2.0 (2.5)	1.1 (1.2)	2.1 (2.4)

Table 3. Absolute errors in isomerization energies (kcal/mol) at HF/6-31++G(d,p) for alpha helixes, relative to fully *ab initio* (no SFM). The non-bonded terms use the combined EFP/*ab initio* approximation (cutoff ≥2.7Å) or *ab initio* (given in parentheses).

Isomer	2nd Order HF/6-31++G(d,p) HF kcal/mol	3rd Order HF/6-31++G(d,p) HF kcal/mol
MAQWUW_1-MAQWUW_2	2.7 (1.1)	1.7 (0.2)
WUYCUO-WUYDAV	5.7 (5.8)	3.9 (6.1)
WUYCUO-WUYDEX	0.9 (2.5)	0.1 (3.1)
YETPES_1-YETPES_2	1.1 (5.3)	0.2 (1.0)
MAE	2.6 (3.7)	1.5 (2.6)

Table 4. Net CPU times (minutes) for the SFM HF/6-31++G(d,p) on a single core of a Xeon 2.66Ghz quad core Cloverton node, with 16 GB RAM. Net times include the time needed for EFP generation. The EFP generation time is given in parentheses. The total number of non-bonded terms is also listed. The heading EFP indicates the use of EFP for the non-bonded terms.

| Isomer | 2nd Oder Non-bonded | | | 3rd Order Non-bonded | | |
	# non-bonded terms	EFP	No EFP	# non-bonded terms	EFP	No EFP
MAQWUW_1	1113	128 (25)	144	3159	329 (191)	559
MAQWUW_2	1113	128 (26)	140	3155	333 (194)	562
WUYCUO	1752	155 (33)	182	5049	413 (214)	853
WUYDAV	1754	162 (33)	188	5059	439 (227)	909
WUYDEX	1754	150 (32)	180	5052	419 (217)	880
YETPES_1	932	115 (24)	122	2623	305 (170)	490
YETPES_2	929	117 (25)	126	2618	299 (166)	497

Table 5. Comparison of the number of *ab initio* non-bonded terms needed for non-bonded EFP/*ab initio* cutoffs set to 2.7 and 4.5Å at the 2nd and 3rd order many body approximation.

| | 2nd Order Many Body | | 3rd Order Many Body | |
	2.7Å (terms)	4.5Å (terms)	2.7Å (terms)	4.5Å (terms)
MAQWUW_1	34	233	113	729
MAQWUW_2	34	224	106	693
WUYCUO	35	318	108	1054
WUYDAV	40	327	130	1075
WUYDEX	36	321	118	1055
YETPES_1	29	225	79	709
YETPES_2	25	225	70	708

Table 6. Absolute errors in the FMO2-MP2 and FMO3-MP2 total energy of the 32 water clusters selected from EFP Monte Carlo/Simulated Annealing simulations. Isomer names are only used to distinguish isomers from one another.

| | Absolute Error (kcal/mol) | | | |
| | 6-31++G(d,p) | | 6-311++G(3df,2p) | |
Isomer[a]	FMO2-MP2	FMO3-MP2	FMO2-MP2	FMO3-MP2
32_1	11.8	2.2	26.8	1.0
32_2	12.4	2.5	28.0	1.2
32_3	11.4	1.9	27.3	1.3
32ab	12.5	2.5	27.4	1.3
32ad	11.8	2.5	27.3	1.2
32h	12.3	2.5	27.3	1.3
32o	12.5	2.5	25.8	1.0
32z	12.4	2.3	24.6	1.2

[a]One water molecule chosen as a monomer

Table 7. Relative FMO2-MP2 and FMO3-MP2 energies of the 32 water clusters selected from EFP Monte Carlo/Simulated Annealing simulations. Isomer names are only used to distinguish isomers from one another.

| | Relative Energies (kcal/mol) | | | | | |
| | 6-31++G(d,p) | | | 6-311++G(3df,2p) | | |
Isomer[a]	FMO2-MP2	FMO3-MP2	*ab initio*	FMO2-MP2	FMO3-MP2	*ab initio*
32_1	0.0	0.0	0.0	0.0	0.0	0.0
32z	1.3	0.5	0.2	0.7	0.2	0.1
32_2	1.4	1.2	0.9	1.1	0.8	0.5
32ab	1.5	1.2	0.9	1.1	0.9	0.6
32h	1.4	1.2	0.9	1.4	1.0	0.7
32o	1.7	1.5	1.2	1.7	1.3	1.0
32ad	4.9	6.0	6.0	5.7	6.1	5.8
32_3	11.8	14.3	14.1	14.0	14.1	14.4

[a]One water molecule chosen as a monomer

Table 8. FMO2 errors (kcal/mol) for tetramer and hexamer ionic liquid clusters.

	Absolute Error (kcal/mol) 6-31+G(d)	
Tetramers	FMO2-MP2	FMO3-MP2
1-H,4-H-1,2,4-triazolium dinitramide	0.06	0.02
1-amino,4-H-1,2,4-triazolium dinitramide	0.69	0.03
Hexamers		
1-H,4-H-1,2,4-triazolium dinitramide	0.32	0.07
1-amino,4-H-1,2,4-triazolium dinitramide	1.35	0.27

Table 9. Comparison of Mulliken charges for all ionic liquid systems investigated.

Tetramers		Mulliken Charges 6-31+G(d)		
		FMO2-MP2	FMO3-MP2	MP2
1-H,4-H-1,2,4-triazolium dinitramide	Cation 1	0.82	0.77	0.74
	Cation 2	0.82	0.77	0.74
	Anion 1	-0.82	-0.77	-0.74
	Anion 2	-0.82	-0.77	-0.74
1-amino,4-H-1,2,4-triazolium dinitramide	Cation 1	0.87	0.84	0.82
	Cation 2	0.82	0.82	0.82
	Anion 1	-0.89	-0.94	-0.93
	Anion 2	-0.80	-0.74	-0.71
Hexamers				
1-H,4-H-1,2,4-triazolium dinitramide	Cation 1	0.86	0.82	0.79
	Cation 2	0.88	0.85	0.80
	Cation 3	0.94	0.91	0.87
	Anion 1	-0.83	-0.79	-0.77
	Anion 2	-0.95	-0.92	-0.88
	Anion 3	-0.90	-0.86	-0.81
1-amino,4-H-1,2,4-triazolium dinitramide	Cation 1	0.84	0.84	0.88
	Cation 2	0.79	0.78	0.76
	Cation 3	0.90	0.89	0.89
	Anion 1	-0.81	-0.92	-0.95
	Anion 2	-0.83	-0.85	-0.83
	Anion 3	-0.88	-0.74	-0.75

Table 10. Timings for ionic liquid clusters performed on a Cray XT4 with 2.1GHz AMD Opteron64 processors. Each node contains a 4 core CPU and 8 GB of RAM.

Tetramer	# CPUs	Timing (minutes) 6-31+G(d) FMO2-MP2	MP2
1-amino,4-H-1,2,4-triazolium dinitramide	8	12.2	28.4
	16	6.4	14.7
	32	3.5	7.3
Hexamer			
1-amino,4-H-1,2,4-triazolium dinitramide	8	24.0	172.1
	16	12.5	83.9
	32	6.8	42.8

Scheme 1. Depiction of isomers used in Table 2. Structures are from the Cambridge Structural Database (CSD). Non-hydrogen atoms have been labeled.
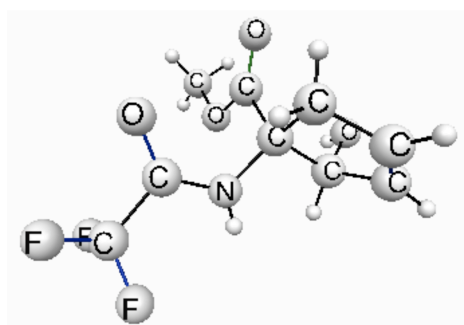
BAZGEP
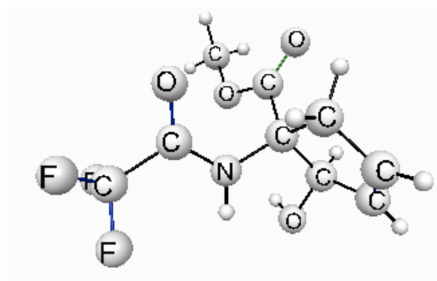
BAZGIT
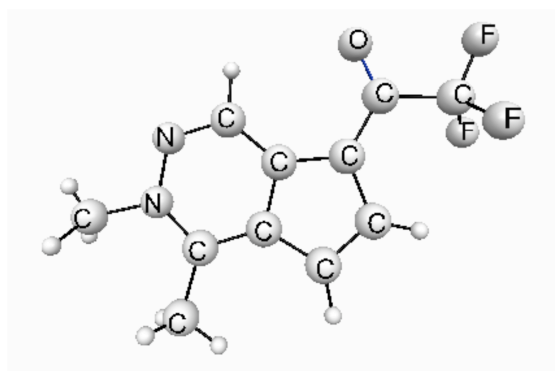
BELDIF
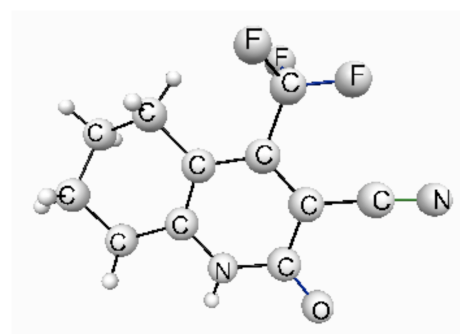
NOTGAE
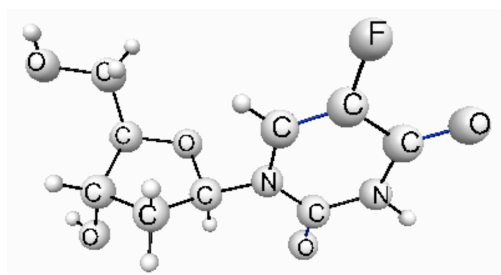
FDOURD01
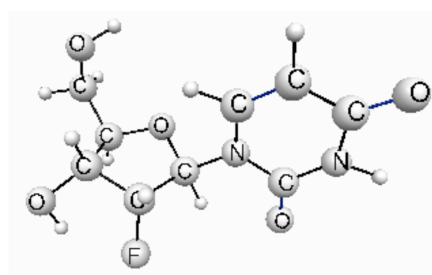
BOFWIC
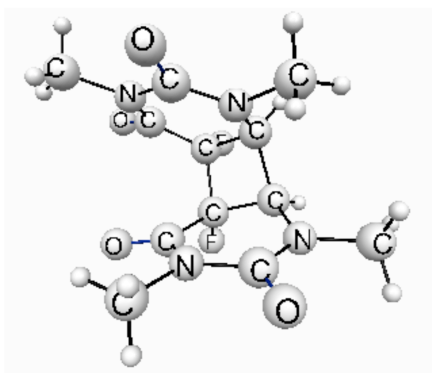
CONBAI
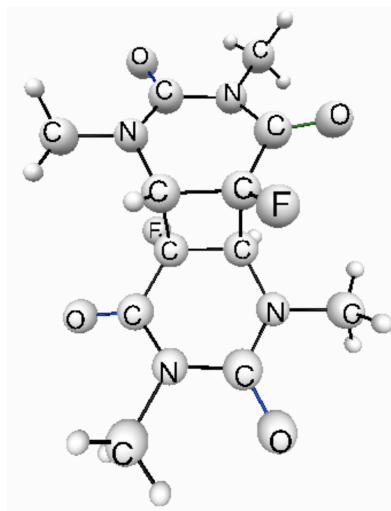
FDMUPDI0
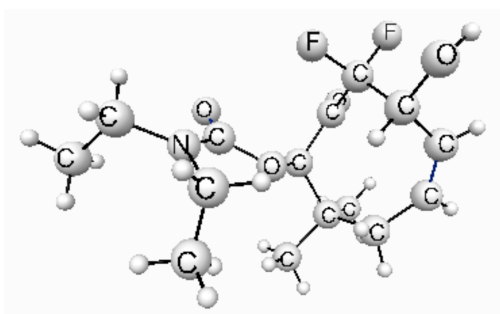
IDUFES
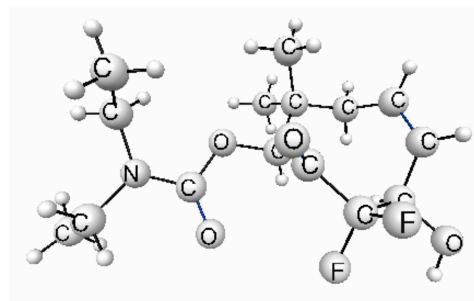
IDUFAO
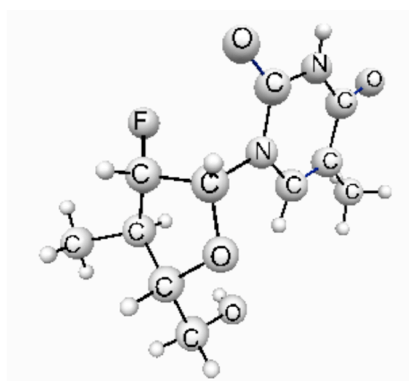
LEDRAN
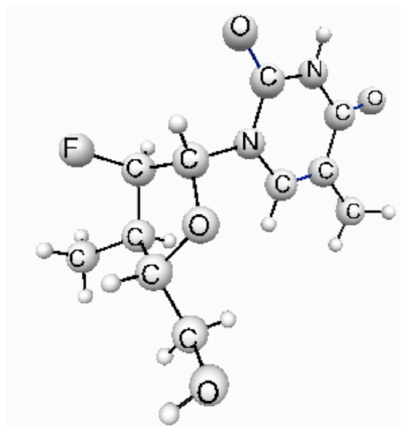
LEDRER
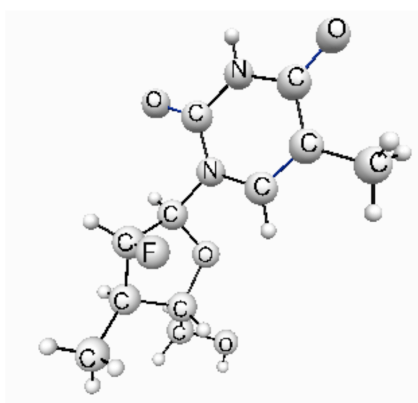
LEDRIV
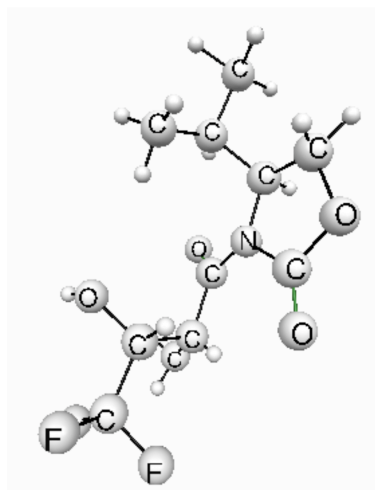
TAXYIA
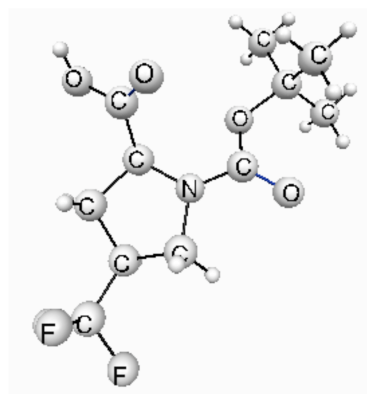
MOGQOO
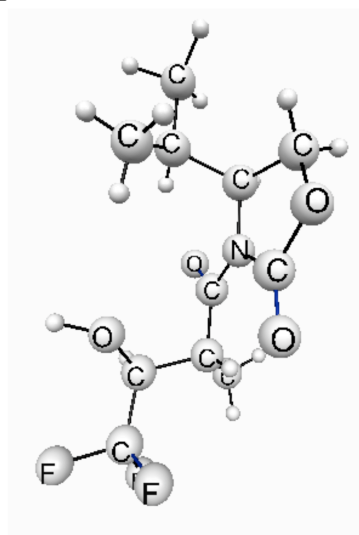
TAXYOG

Scheme 2. Depiction of alpha helix isomers used in Table 4. Structures are from the Cambridge Structural Database (CSD). Non-hydrogen atoms have been labeled.

MAQWUW_1

MAQWUW_2

WUYCUO

WUYDAV

WUYDEX
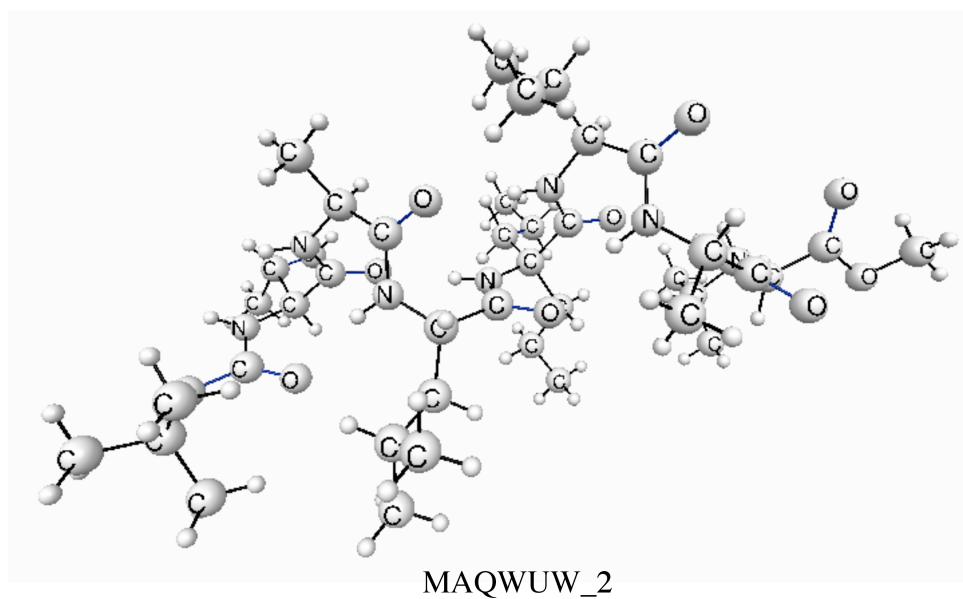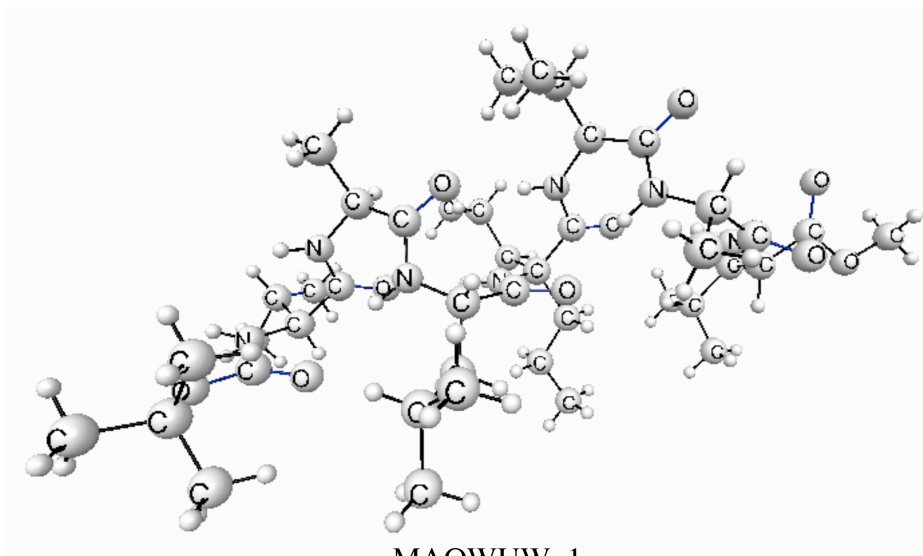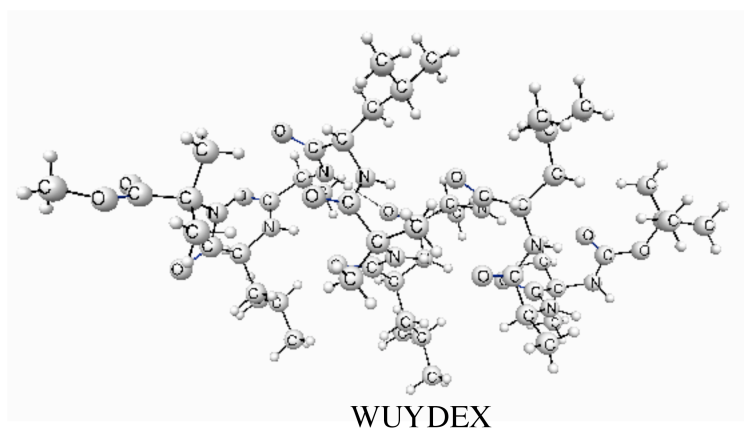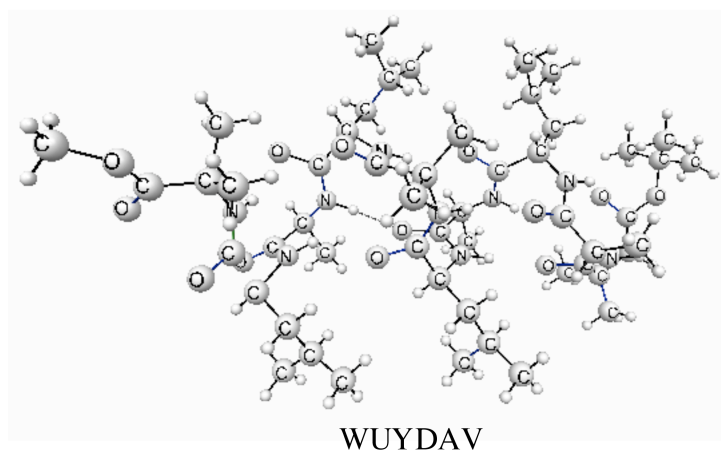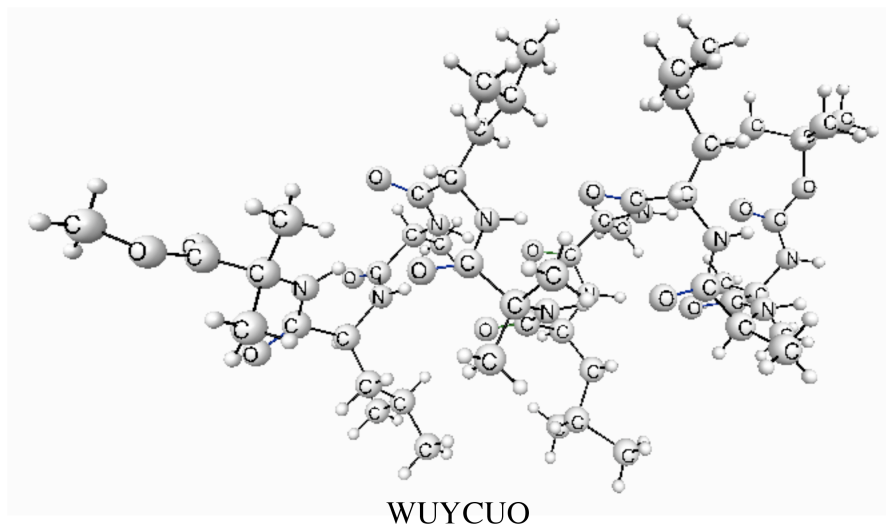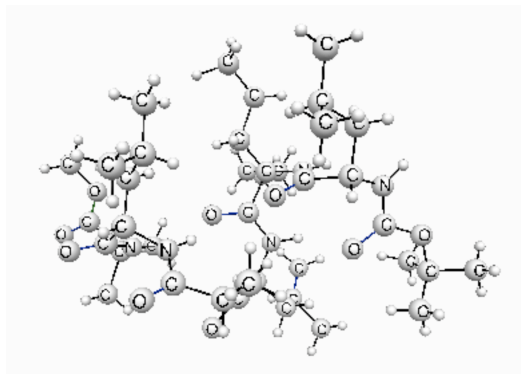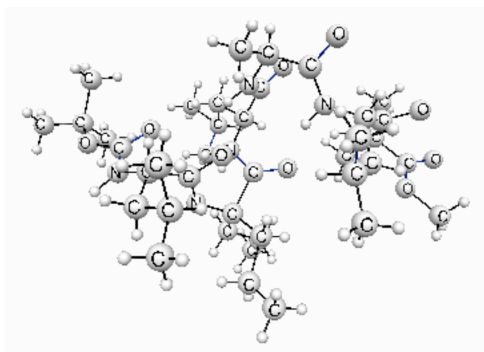
YETPES_1

YETPES_2

FIGURE CAPTIONS

Figure 1. Sandwich, T-shaped, and parallel-displaced configurations of the benzene dimer.

Figure 2. Comparison of SAPT and EFP interaction energy (kcal/mol) decomposition as a function of the seperation (Å) of benzene dimer in the sandwhich configuration.

Figure 3. Comparison of SAPT and EFP interaction energy (kcal/mol) decomposition as a function of the seperation R (Å) of benzene dimers in the T-shaped configuration.

Figure 4. Electrostatic fractioning of bonds.

Figure 5. Monomer calculation performed in the ESP of the full system.

Figure 6. Illustration of FMO approximations applied to three monomers I, J, L (left) and as applied to dimer IJ and monomer L (right).

Figure 7. Lowest energy cluster of 32 water molecules obtained from EFP Monte Carlo/Simulated Annealing simulations.

Figure 8. Ion pairs of 1-amino,4-H-1,2,4-triazolium dinitramide (top) and 1-H,4-H-1,2,4-triazolium dinitramide (bottom).

Figure 9. Lowest energy structure of 1-amino,4-H-1,2,4-triazolium dinitramide tetramer obtained from an *ab initio* MP2/6-31+G(d) optimization.

Figure 10. Lowest energy structure of 1-amino,4-H-1,2,4-triazolium dinitramide hexamer obtained from an *ab initio* MP2/6-31+G(d) optimization.

Figure 11. Lowest energy structure of 1-H,4-H-1,2,4-triazolium dinitramide tetramer obtained from an *ab initio* MP2/6-31+G(d) optimization.

Figure 12. Lowest energy structure of 1-H,4-H-1,2,4-triazolium dinitramide hexamer obtained from an *ab initio* MP2/6-31+G(d) optimization.

Figure 1.

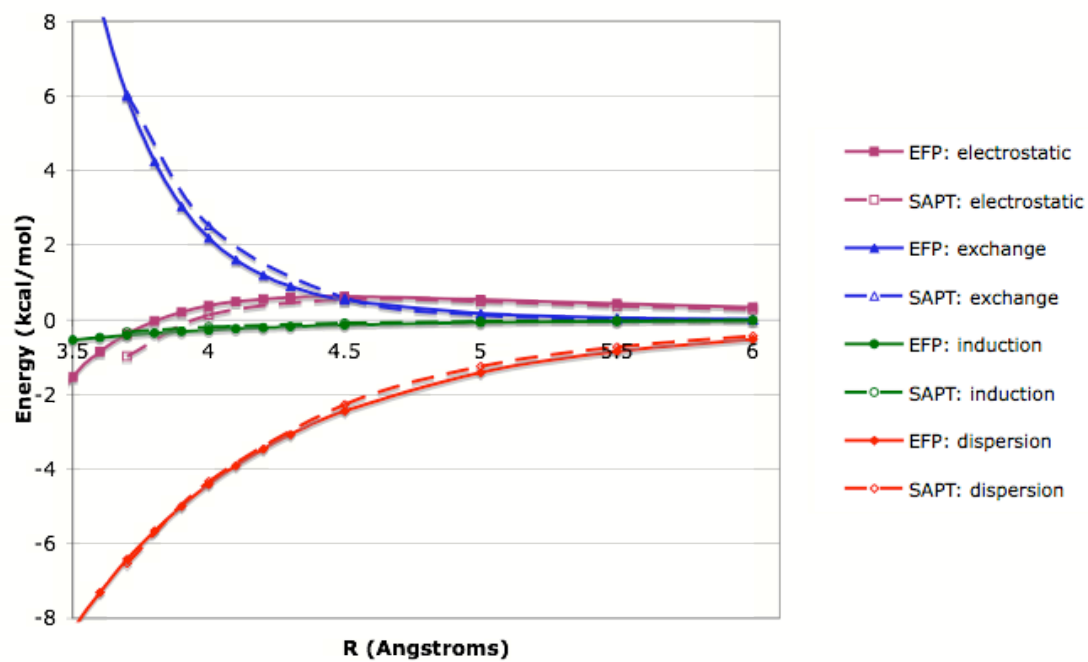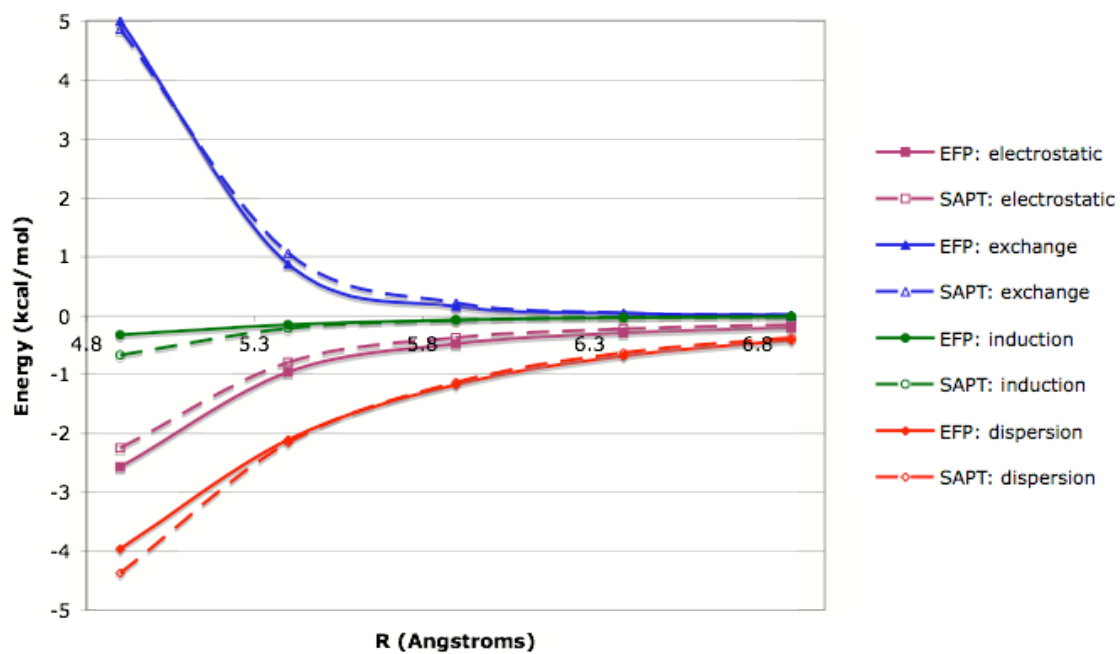sandwich   T-shaped   parallel-displaced
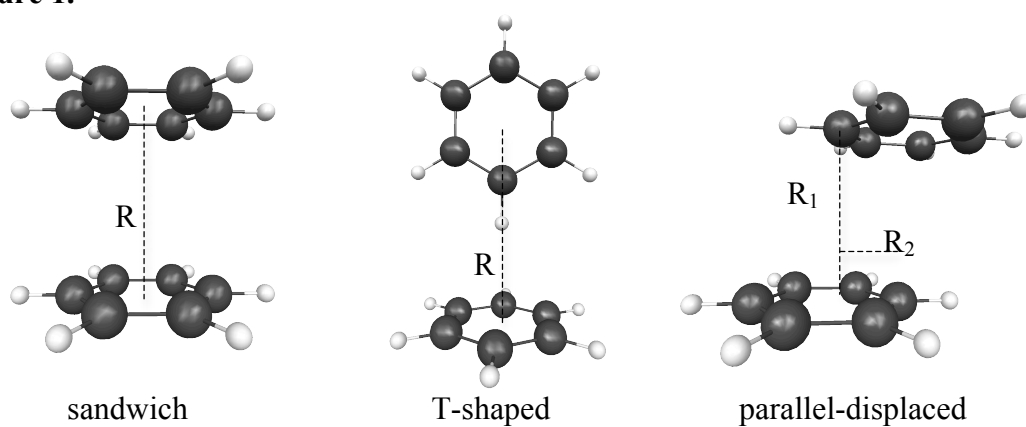
Figure 2.

Figure 3.

Figure 4.

Figure 5.

Coulomb "bath" of the full system being taken into account

monomer electron density being calculated

× = monomer

Figure 6.

I, J and L all as
separate monomers

I and J composing dimer IJ

L as a monomer

- – – – Monomer I approximation cut-off
- .... Monomer J approximation cut-off
- ——— Monomer L approximation cut-off
- ——— Dimer IJ
- – – – Dimer IJ approximation cut-off
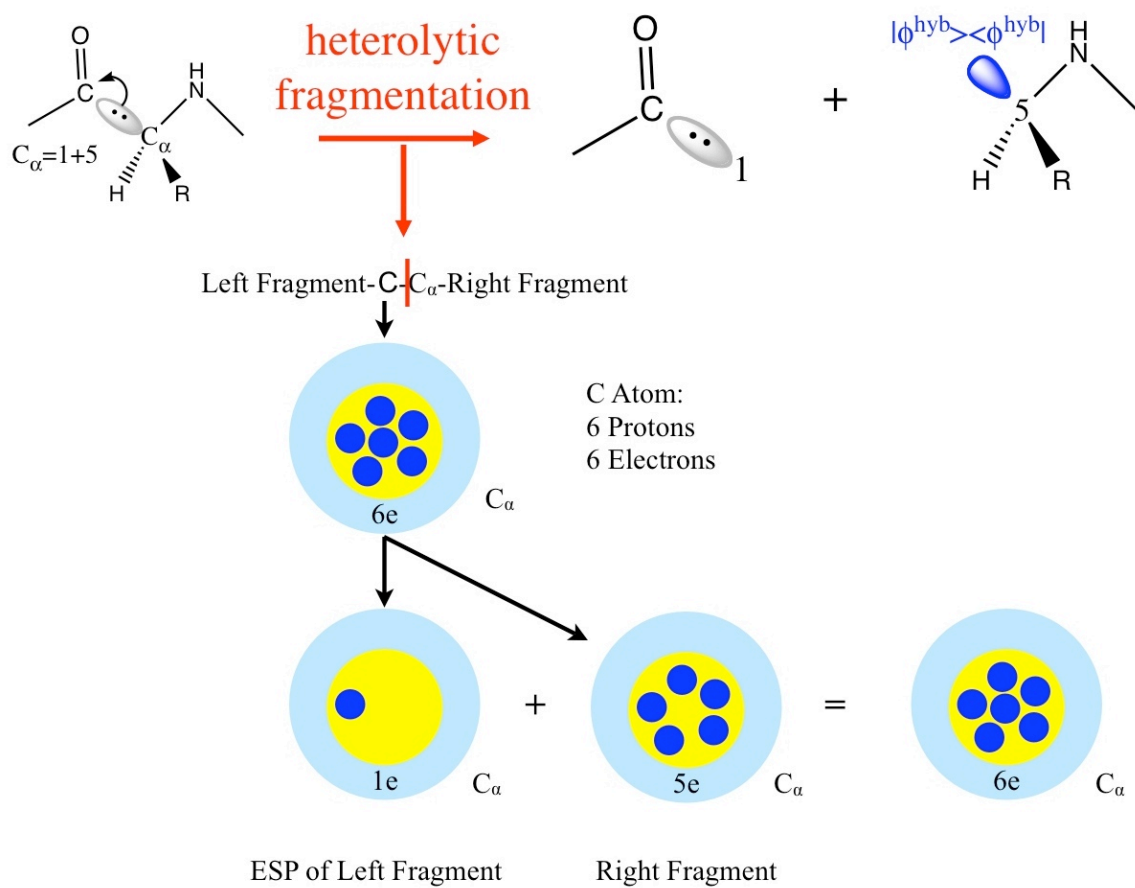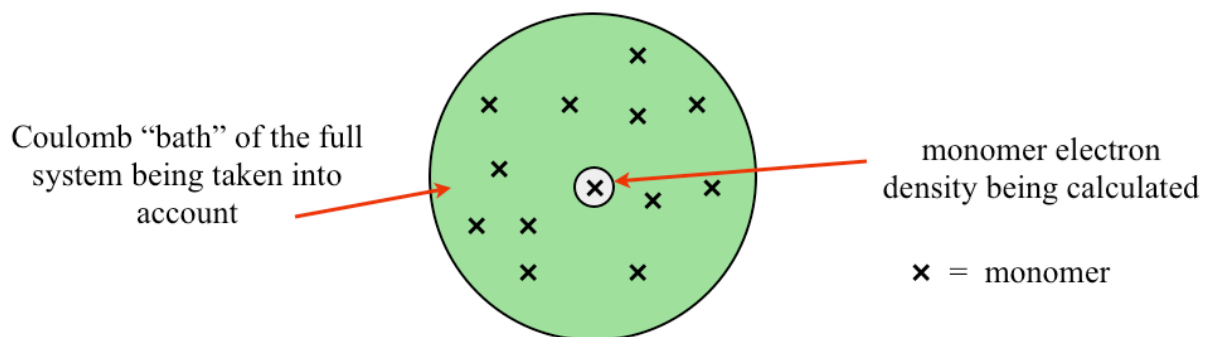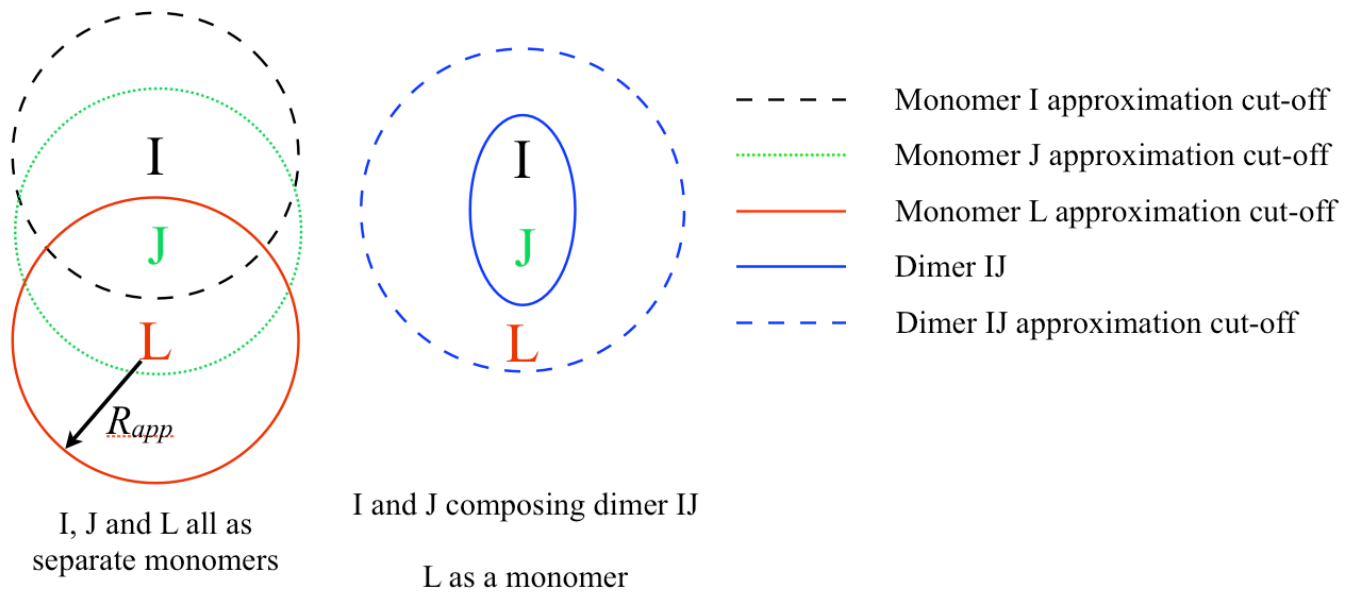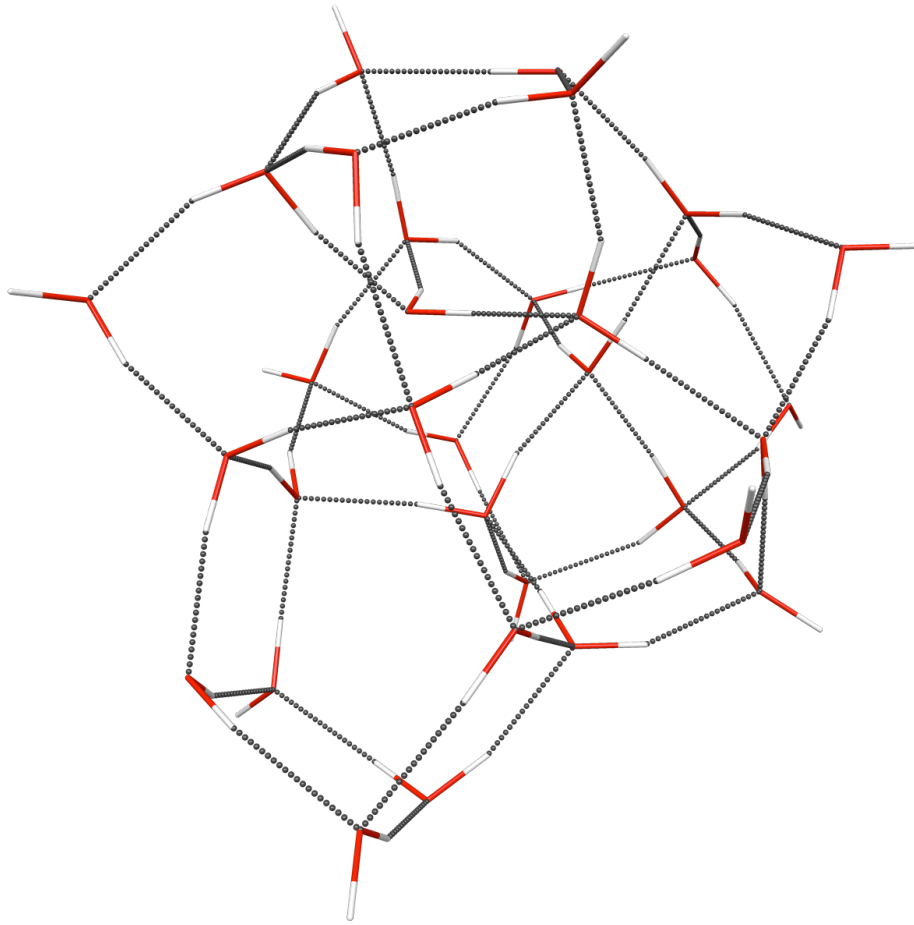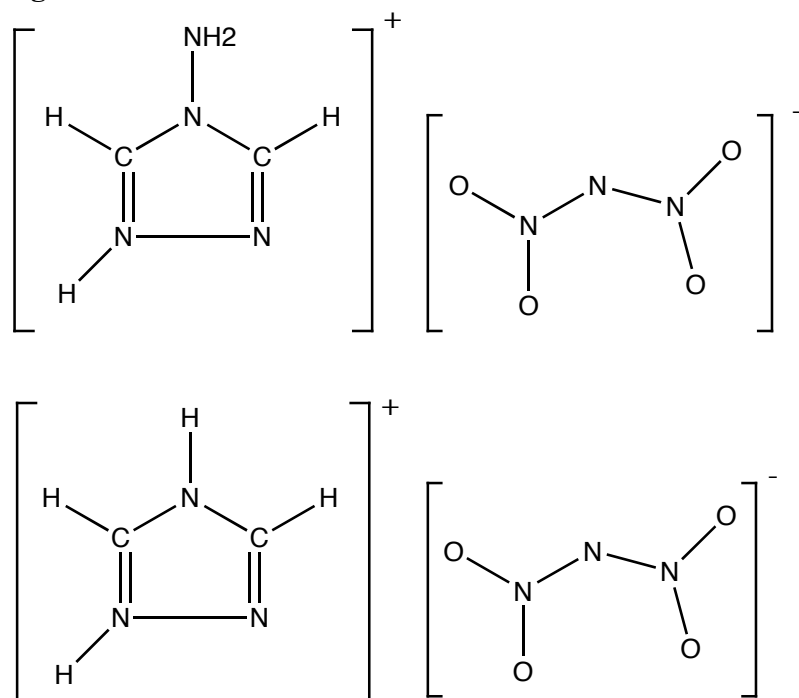
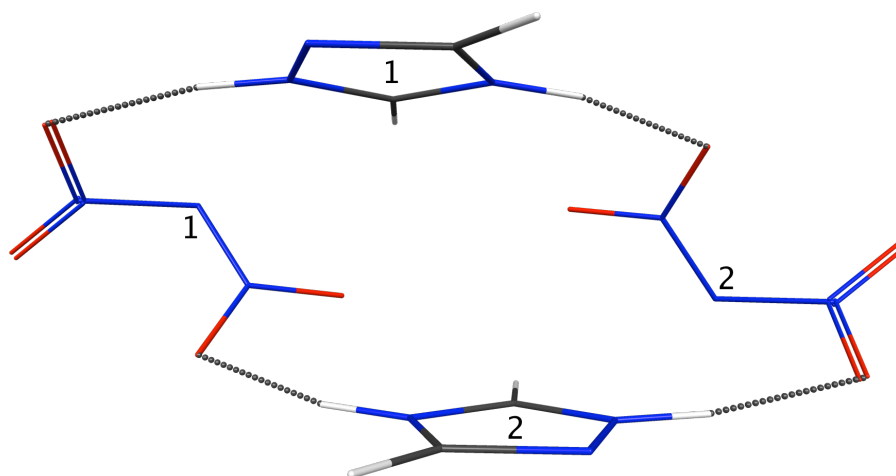**Figure 7.**

Figure 8.

Figure 9.
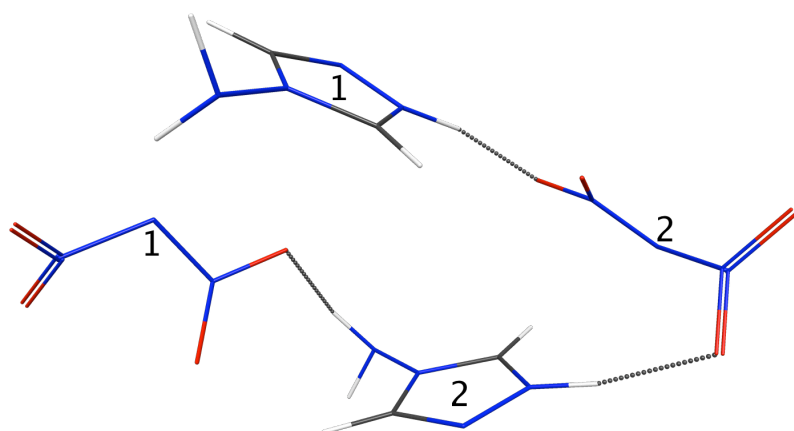
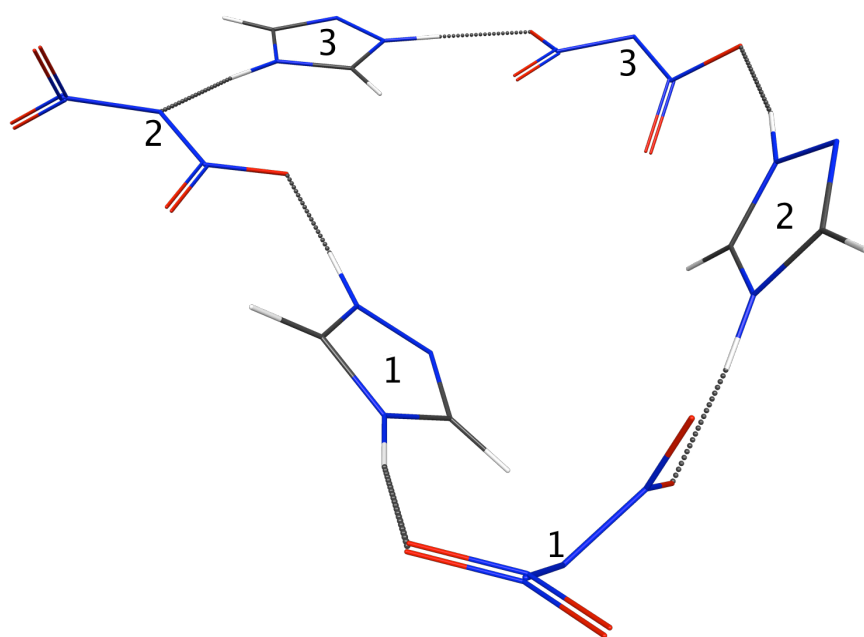**Figure 10.**
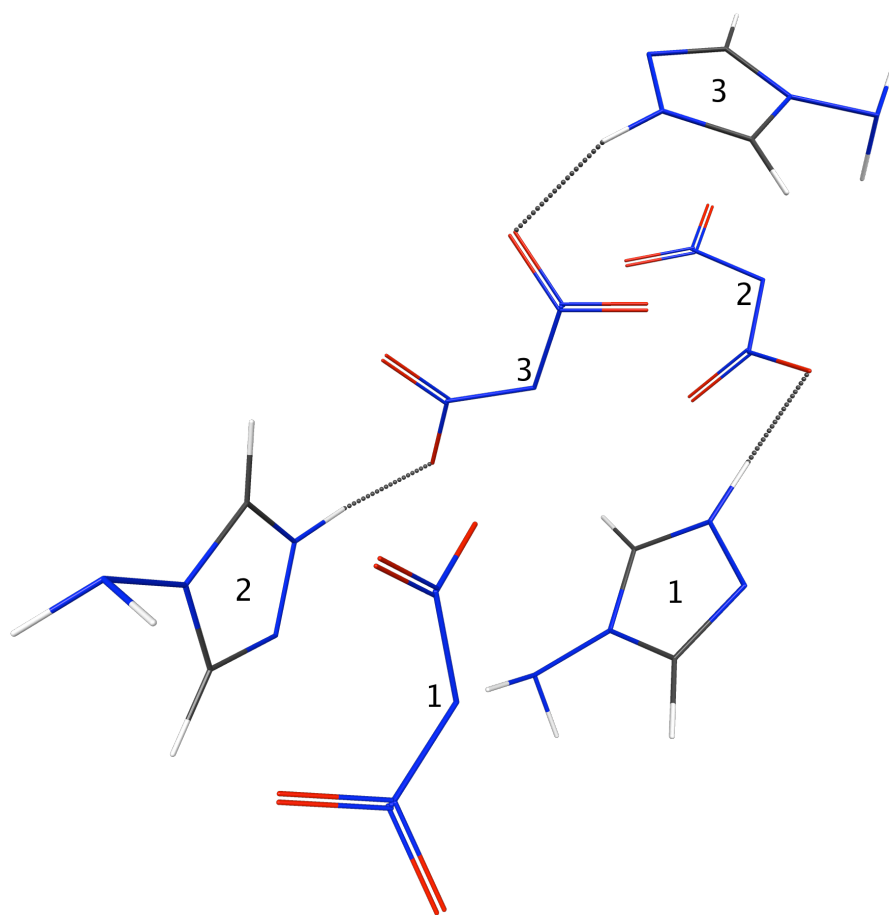
Figure 11.

Figure 12.

SYSTEMATIC FRAGMENTATION METHOD AND THE EFFECTIVE FRAGMENT POTENTIAL: AN EFFICIENT METHOD FOR CAPTURING MOLECULAR ENERGIES

Jonathan M. Mullin,[§] Luke B. Roskop,[§] Michael A. Collins,* and Mark S. Gordon[§].

[§]Iowa State University, Ames Laboratory, Ames, IA, USA

*Australian National University, Canberra, ACT, AU

Abstract

The systematic fragmentation method has been shown to accurately approximate full *ab initio* calculations. Sub kcal/mol accuracy requires including non-bonded interactions; these are computationally demanding due to the sheer number of interactions. A simple approximation to the non-bonded interactions using classical electrostatics can work reasonably well in some cases, but is not a general solution. Comparatively, the effective fragment potential method provides a generally useful representation of non-bonded interactions while still providing an approach that is much more computationally efficient than *ab initio* calculations.

1. Introduction

Theoretical chemistry has progressed to the point that it is able to achieve "chemical accuracy" (~1 kcal/mol) for predicting the energetics related to chemical processes for small molecules and to approach chemical accuracy for moderately sized molecules. The ability to predict structures, energies, and dynamics for chemical processes can be accomplished through *ab initio* quantum mechanics (QM) methods. The most commonly used QM methods include density functional theory (DFT)[1] and Hartree-Fock (HF)[2]. HF is generally used as the starting point for more accurate approaches that include electron correlation, such as second order perturbation theory (MP2)[3] and coupled cluster theory with single, double and perturbative triple excitations (CCSD(T))[4]. For systems requiring more than a single electronic configuration, multi-reference methods such as multi-configurational self-consistent field (MCSCF)[5] multi-reference perturbation theory (MRPT)[6] and multi-reference configuration interaction (MRCI)[7] are used. All of these correlated methods are computationally demanding, scaling as N^5 or worse where N is the number of basis functions. To achieve chemical accuracy, one needs to combine these methods with large atomic basis sets[8] exacerbating computational demands. In order to extend these highly accurate methods to significantly larger chemical systems, one needs to develop methods with better scaling. There have been many approaches to accomplishing this, including the use of localized molecular orbital (LMO) methods,[7-12] fast multipole methods (FMM),[13-17] and fragmentation methods in which a large molecular system is fragmented into smaller, more tractable pieces.

Several general fragmentation methods have been proposed, including molecular fragmentation with conjugated caps (MFCC),[18] the elongation method,[19] the molecular tailoring

approach (MTA),[20] the fast electron correlation method for molecular clusters developed by Hirata,[21] Truhlar's electrostatically embedded many-body (EE-MB) expansion,[22] multi-centered QM/QM methods,[23] the fragment molecular orbital (FMO) method[24] and the systematic fragmentation method (SFM).[25] Initial fragmentation models[26] built large molecules with a set of fragments from common chemical groups (carboxyl, amino, etc.). Newer fragmentation schemes begin with the un-fragmented system and break it into smaller parts (fragments). Fragmentation schemes need to account for the chemical environment of the full system when calculating each fragment. The chemical environment is usually treated in an approximate manner to retain the effect of the full system. SFM and its approach to capturing the chemical environment are considered in this work.

2. Systematic Fragmentation Method

In order to study large molecular systems, the systematic fragmentation method (SFM) considers several sub-systems, or "groups", independently. By treating each set of overlapping groups with accurate levels of theory, the total energy of the system, and other properties, may be obtained by addition and subtraction of the contributions from individual groups. This allows SFM to achieve a significant decrease in computational expense while retaining good accuracy when compared to full *ab initio* calculations. A complete description of the energy of the system is recovered via a perturbative many-body non-bonded term. In the original SFM formulation[25] this non-bonded term was obtained using a classical electrostatic potential. SFM is formally independent of the electronic structure method and the atomic basis set. Scaling approaches linearity as the system size is increased.

2.1 Background

In general, molecules can be considered to be composed of sets functional groups that are bonded to each other. In this sense, the "groups" that are employed in the SFM can be thought of as functional groups. To illustrate the SFM consider ethanol, which contains three functional groups (CH_3, CH_2, and OH). Figure 1 illustrates the fragmentation of ethanol into its component substituent functional groups, by breaking two single bonds. In each case, the bonding pair of electrons is split, each fragment being assigned one electron from the original bonded pair. A "cap" (hydrogen atom) is applied to the dangling bonds that are created by the fragmentation to avoid the creation of radical species. Double or triple bonds are not broken, retaining the relevant atoms as a part of one functional group. A second example in Figure 1 is ethanal which contains two functional groups (CH_3 and CHO). The hydrogen capping yields three molecules, CH_4, CH_4, and H_2O for ethanol, and two molecules, CH_4 and CH_2O, for ethanal.

The SFM may be implemented at several levels of theory; these levels are defined as follows.[25b]

Consider the molecule *M*:

$$M = G_1G_2G_3G_4G_5G_6G_7G_8 \tag{1}$$

At SFM level 1, each broken bond is separated by one functional group G_i. The initial fragmentation would create the following pieces:

$$M \approx G_1G_2 + G_2G_3G_4G_5G_6G_7G_8 - G_2 \tag{2}$$

G_2 is subtracted in order to conserve the number of atoms. The process of fragmentation is repeated on the $G_2G_3G_4G_5G_6G_7G_8$ fragment until no fragment remains that is larger than two functional groups. The energy of M, E(M), can be approximated by a simple sum of fragment energies. For SFM level 1 this is:

$$E_{\text{level 1}}^{\text{bonded}}(M) = E(G_1G_2) + E(G_2G_3) + E(G_3G_4) + E(G_4G_5) + E(G_5G_6) + E(G_6G_7) + E(G_7G_8)$$
$$- E(G_2) \quad - E(G_3) \quad - E(G_4) \quad - E(G_5) \quad - E(G_6) \quad - E(G_7) \tag{3}$$

The superscript "bonded" means that non-bonded interactions are not included in E(M) at this point.

At SFM level 2, each broken bond is separated by two functional groups. The energy of *M* is approximated by the following expression:

$$E_{\text{level 2}}^{\text{bonded}}(M) = E(G_1G_2G_3) + E(G_2G_3G_4) + E(G_3G_4G_5) + E(G_4G_5G_6) + E(G_5G_6G_7) + E(G_6G_7G_8)$$
$$- E(G_2G_3) \quad - E(G_3G_4) \quad - E(G_4G_5) \quad - E(G_5G_6) \quad - E(G_6G_7) \tag{4}$$

At SFM level 3, each broken bond is separated by three functional groups. The energy of *M* is approximated by the following expression:

$$E_{\text{level 3}}^{\text{bonded}}(M) = E(G_1G_2G_3G_4) + E(G_2G_3G_4G_5) + E(G_3G_4G_5G_6) + E(G_4G_5G_6G_7) + E(G_5G_6G_7G_8)$$
$$- E(G_2G_3G_4) \quad - E(G_3G_4G_5) \quad - E(G_4G_5G_6) \quad - E(G_5G_6G_7) \tag{5}$$

SFM has two formal limitations; the first is that if there is conjugation in delocalized molecular systems (e.g., butadiene), the entire delocalized moiety must remain intact (not fragmented). The second limitation is that one cannot fragment six-member rings using level 3, because capping hydrogens approach each other too closely. This causes unphysical repulsive interactions. SFM level 2 cannot fragment five member rings, while four and three member rings cannot be fragmented at all. A ring repair rule requires the ring to remain intact and to be considered a functional group itself.

2.2. Non-bonded interactions

A simple sum of the individual hydrogen-capped fragment energies neglects interactions among the separated (non-bonded) fragments. In the full *ab initio* calculation these non-bonded

interactions are inherently part of the calculation. By assuming that bonded interactions are greater in strength than non-bonded interactions, a modified many body expansion may be employed to model the non-bonded interactions.

2.2.1 Two-body interactions

Two arbitrary functional groups G_1 and G_2 are placed in their positions in the original full molecule M, and their interaction energy is given by

$$E_{nb}^{(1,1)}[G_1;G_2] = E(G_1G_2) - E(G_1) - E(G_2), \qquad (6)$$

where $E(G_1), E(G_2)$ are the one-body fragment energies and $E(G_1G_2)$ is the super-molecular energy of the two separated functional groups. All possible pairs of functional groups (not considered in the bonded calculation) are considered in the total two-body non-bonded energy of the system. Therefore, all pairs of arbitrary functional groups (G_1, G_2) not contained in any one fragment are considered.

2.2.2 Three-body interactions

Mutual interactions among three arbitrary functional groups G_1, G_2 and G_3 are assumed to be negligible unless two of the groups are bonded to one another. Considering G_3 bonded to G_2 yields the three-body interaction energy:

$$\begin{aligned} E_{nb}^{(1,2)}[G_1;G_2,G_3] = {} & E(G_1G_2G_3) - E(G_1) - E(G_2G_3) \\ & - E_{nb}^{(1,1)}[G_1;G_2] - E_{nb}^{(1,1)}[G_1;G_3] \end{aligned}$$

$$. \qquad (7)$$

In Eq. (7) the three-body energy is the super-molecular energy, $E(G_1G_2G_3)$, minus the one-body energies ($E(G_1)$, $E(G_2G_3)$) minus the two body non-bonded energy terms containing any group (G_1) with either of the bonded functional groups (G_2 or G_3). This scheme can be extended to higher order non-bonded interactions. Terms up to n^{th} order form the modified many-body

approximation. This work only considers two and three body interactions.

Addition of the bonded and non-bonded terms yields the total SFM energy,

$$E_{SFM}^{total} = E^{bonded} + E^{non\text{-}bonded} .$$ (8)

Previous work[25c,d] has shown that level 3 with non-bonded interactions is needed to achieve high accuracy (1 kcal/mol) compared to fully *ab initio* calculations.

2.3. Effective Fragment Potential method

The generalized effective fragment potential (EFP2) method[27] is a first-principles based model potential for the evaluation of intermolecular forces. EFP2 is an extension of the EFP1 water model to general systems.[28-31] The EFP-EFP interaction consists of five terms for the EFP2 model potential:

$$E = E_{coul} + E_{ind} + E_{exrep} + E_{disp} + E_{ct} .$$ (9)

These terms correspond, respectively, to Coulombic (electrostatic), induction (polarization), exchange repulsion, dispersion (van der Waals), and charge transfer. Damping functions are employed for the Coulomb, induction and dispersion terms, to account for short- or long-range behavior. Each of these terms may be thought of as a truncated expansion.

The EFP interactions decay either exponentially with respect to distance for short range interactions or as $(1/R)^n$ for long range interactions. The Coulombic, induction, and dispersion are long-range interactions. The exchange repulsion and charge transfer are short-range interactions. A more complete description of EFP has been detailed previously;[27-31] a short overview of the terms is outlined below.

The Coulomb interaction is obtained via the Stone distributed multipolar analysis.[32] This expansion is truncated at the octopole term. Atom centers and the bond midpoints are used as expansion points.

Induction (polarization) is the interaction of an induced dipole on one fragment with the permanent dipole on another fragment, expressed in terms of the dipole polarizability. Truncating at the first (dipole) term is viable, since the molecular polarizability tensor is expressed as a (distributed) tensor sum of localized molecular orbital[33] (LMO) polarizabilities. The number of polarizability points is equal to the number of bonds and lone pairs in the system.

An expansion in the intermolecular overlap integral, in a frozen LMO basis, is used to evaluate the exchange repulsion.[34] The expansion in terms of frozen LMOs on each fragment allows truncation at the quadratic term. Since the basis set required to compute the exchange repulsion interaction is used only to calculate overlap integrals, the computation is very fast and quite large basis sets are realistic.

Dispersion interactions are often expressed by an inverse R expansion,

$$E_{disp} = \sum_{n} C_{n} R^{-n}$$
(10)

The coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities integrated over the entire frequency range.[35] The first term in the expansion, n=6, corresponds to the induced dipole-induced dipole (van der Waals) interactions. In the EFP2 method, this term is derived from the time-dependent HF method. Additionally the contribution of the n=8 term is estimated. The C_6 coefficients are derived in terms of interactions between pairs of LMOs on the two interacting fragments.

Charge transfer interactions in the EFP2 method are derived using a second order perturbative treatment of the intermolecular interactions for a pair of molecules at the Hartree–Fock level of theory.[36] The energy is evaluated by determining the energy due to the occupied valence molecular orbitals on one fragment interacting with the virtual orbitals on another fragment. Charge transfer can be important for polar or ionic species.

The use of classical approximations for the Coulomb and induction terms requires short-range interactions to be moderated by an additional damping function. A classical (point multipole) Coulombic interaction becomes too repulsive at short range; this can be corrected by the damping term.[37] Induction, on the other hand, becomes too attractive at short distances, so a damping term is employed here as well. The form of these damping functions is an exponential which augments the electrostatic multipole of the form:

$$f_{\text{damp}} = 1 - \exp(-\alpha R), \tag{11}$$

where parameters α are determined at each multipole expansion point by fitting the damped multipole potential to the Hartree–Fock potential. Damping terms for the electrostatic and induction terms are derived explicitly from the damped potential and the charge density. Short-range dispersion interactions should decrease to zero. Therefore, each dispersion term is multiplied by a damping function as well. In the work described here, two damping methods are explored, one due to Tang and Toennies[38] and a new approach based on the intermolecular overlap.[39]

The five EFP interaction terms have different relative costs. Considering small EFP clusters, and considering the Coulomb term to be one time unit, the induction interaction would cost approximately one unit, dispersion one unit, exchange repulsion would cost five units, and charge transfer ~50 units. Thus some trade off between computation time and completeness of the potential may be considered. EFP calculations are basis set dependent. The smallest recommended basis set is 6-31++G(d,p).[40] The computational cost of an EFP calculation appears primarily in the initial generation of the EFP, not in EFP-EFP interactions. Therefore, one can employ much larger basis sets with minimal cost. In the tests presented below on the SFM method, the 6-311++G(3df,2p)[41] basis is used. EFP contains no empirically fitted parameters,

allowing automatic creation by a "makefp" run in the GAMESS[42] suite of programs. It has been demonstrated that the EFP2-EFP2 intermolecular (e.g., non-bonded) interaction energies are frequently as accurate as MP2, at a small fraction of the MP2 cost. This suggests the efficacy of determining the SFM non-bonded interactions using this method.

3. Computational details

ZOVGAS (Figure 2) is a compound selected from the Cambridge Structural Database(CSD)[43] containing a combination of several first and second row atoms. So, it provides a nice test of the SFM. Comparisons of the performance in determining the SFM non-bonded contribution are demonstrated by calculating the non-bonded interactions via the CCSD(T)[4] and EFP methods, using the 6-311++G(3df,2p) basis set.

Retinal structures (Figure 3) from the CSD (TRETAL02, CRETAL01) were optimized with MP2/6-311G(3df,2p) and verified by Hessians to confirm the stationary points. The SFM bonded energy is treated at level 3 fragmentation using MP2/6-311++G(3df,2p) and EFP/6-311++G(3df,2p) for the non-bonded level 3 term. The alpha helix structures (MAQWUW, WUYCUO, WUYDAV, WUYDEZ, YETPES) shown in Figure 4 are taken from the CSD. HF single points are examined using the 6-31++G(d,p) basis set. MP2 calculations on the full system would have been prohibitively expensive. HF/6-31G++G(d,p) SFM bonded energies are determined at level 3 fragmentation, with EFP/6-311++G(3df,2p) non-bonded interactions performed with 2-body and 3-body interactions.

The EFP dispersion term has been treated by overlap damping unless otherwise noted. All EFP terms are used including charge transfer for all non-bonded interactions. All calculations were carried out with the GAMESS suite of programs.[42]

4. Results

4.1 *ZOVGAS*

The non-bonded energy of ZOVGAS (Figure 2) consists of 129 separate interactions, including 1-, 2-, and 3-body terms. The shortest non-bonded atom-atom distances range from 2.3 - 4.6 Å. Table 1 shows these non-bonded interaction energies calculated using MP2, CCSD(T), and EFP. The two EFP columns in this table differ only in the method that was used for dispersion damping: Tang-Toennies damping in column 3 of the table and the new overlap damping in column 4. The MP2 and CCSD(T) non-bonded interaction energies are in good agreement with each other, while the EFP method overestimates the non-bonded interaction energy. It is clear from the last two columns of the table that the overlap dispersion damping is more reliable, as suggested in Ref. 39. In Figure 5 the difference between the EFP and CCSD(T) interaction energies is plotted vs. the nearest atom-atom distance for each non-bonded pair. As the nearest atom-atom distance decreases, the EFP error relative to CCSD(T) systematically increases. In general, there is good agreement between EFP and CCSD(T) for non-bonded distances that are larger than 2.7 Å. For longer non-bonded atom-atom distances, the EFP error (Table 1) is small ($\sim \pm 0.1$ mh), and the two damping methods are in good agreement with each other. Therefore, in subsequent SFM calculations, the QM method of choice (e.g., CCSD(T)) is used for non-bonded distances that are 2.7 Å or shorter, while EFP is used for distances longer than 2.7 Å. This is monitored by using a distance cutoff d_{min}. The use of the overlap dispersion damping is recommended and used in the remainder of the study.

It is interesting to consider the basis set superposition error (BSSE) for the CCSD(T) non-bonded interactions. Since EFP is a model potential, it has no BSSE, while the quantum methods do. Ten of the ZOVGAS non-bonded interactions, with atom-atom distances ranging from 2.3 -

4.4 Å were used to calculate the CCSD(T) BSSE. This was done using the standard counterpoise method,[44] although Truhlar has noted that such corrections can be unreliable.[45] For these 10 cases, the CCSD(T) BSSE is of the same order of magnitude as the interaction energy itself. So, the EFP errors may be less than they appear in Table 1.

The origin of the errors in the non-bonded interactions can be assessed by comparing the EFP-EFP interactions with those obtained from symmetry adapted perturbation theory (SAPT)[46]. For the same set of 10 test interactions noted in the previous paragraph, EFP and SAPT are in good agreement, except for the Tang-Toennies damped dispersion interactions where the atom-atom distances are less than 3 Å. The Tang-Toennies damping with a fixed parameter in the exponent[33a] may not be adequate to correctly capture the dispersion interactions at short range. Nonetheless, the agreement between CCSD(T) and EFP used only for distances greater than 2.7 Å is excellent.

Table 2 compares the CPU times required to calculate the 129 non-bonded interactions with CCSD(T), with EFP, and with EFP only for non-bonded interactions whose atom-atom distances are greater than 2.7 Å (called "mix" in the table). The CCSD(T) calculations required more than 5,000 CPU hours, whereas the EFP calculations required ~1 minute. Creation of the potentials themselves took just over four hours for all 37 unique fragments. The "mix" approach, which is in excellent agreement with CCSD(T), required ~ 500 CPU hours, an order of magnitude less than CCSD(T) itself.

4.2 *Retinal*

Retinal is a well-known example of a cis-trans isomerization. One of the goals in any model chemistry methods is the ability to reproduce relative energies of a system. Using d_{min} = 2.7 Å, the *cis-trans* isomerization of retinal (Figure 3) was compared for fully *ab inito* MP2/6-

311++G(3df,2p) and SFM. The SFM calculations employ EFP for the non-bonded interactions with d_{min} = 2.7. Table 3 shows that SFM is able to reproduce the MP2 relative energies to within 0.5 kcal/mol of the full *ab inito* calculations.

4.3 *Alpha Helix*

The motivation to use the EFP method for non-bonded interactions, rather than a point charge model, was based on the fact that electrostatic treatment of the non-bonded term does not capture many body polarization effects such as the stacked dipole of an alpha helix. As an example WUYDAV (Figure 4) has a dipole moment of 55.6 D. Table 4, shows the SFM and HF/6-31++G(d,p) isomer energies for model alpha helices, ranging from 125-170 atoms. Increasing the non-bonded term from level 2 to level 3 improves the mean absolute error by ~1 kcal/mol, while on average doubling the time required to calculate the non-bonded interactions. The SFM errors obtained using the overlap dispersion damping with d_{min} =2.7Å are ~ 1 kcal/mol smaller than when the non-bonded term is treated via *ab inito* or EFP alone. This suggests that the EFP calculations may, in fact, be more reliable than HF for the non-bonded interactions.

5. Conclusions

The non-bonded interactions in the three systems considered here explore a broad range of such interactions that are encountered in organic molecular systems. It is apparent from the results presented here that the EFP method presents a viable way to determine non-bonded interactions in fragmentation methods like the SFM for nearest atom-atom distances larger than 2.7 Å. At shorter non-bonded distances, the EFP method is not as reliable, especially when fixed-parameter Tang-Toennies dispersion damping is used. However, it appears that the CCSD(T) BSSE is significant, and this would ameliorate the apparent EFP error. The dispersion overlap significantly improves the performance of the EFP approach. The prescription presented here, to

use the *ab initio* method of choice for interactions with atom-atom distances ≤ 2.7 Å, yields ~kcal/mol accuracy with an order of magnitude improvement in the computational cost. Using the EFP for all non-bonded terms achieves three orders of magnitude improvement in the computational cost, but does not generally yield kcal/mol accuracy.

Acknowledgements

References

1. Hohenberg, P.; Kohn, W. *Phys. Rev*. **1964**, *B864*, 136
2. (a) Hartree, D.R. *Proc. Cambridge Phil. Soc*. **1928**, *24*, 426. (b) Hartree, D. R. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 111. (c) Hartree, D. R. *Proc. Cambridge Phil. Soc.* **1928**, *24*, 89. (d) Fock, V. *Physik.* **1930**, *61*, 126. (d) Roothan, C. C. *J. Rev. Mod. Phys.* **1960**, *32*, 179.
3. Møller, C. Plesset, S. Phys. Rev. **1934**, *46*, 618.
4. (a) Paldus, J., Wilson, S., editor. Handbook of Molecular Physics and Quantum Chemistry, vol. 2. Chichester: Wiley; **2000**; p. 272–313. (b) Cízek, J. *J. Chem. Phys.* **1966**, *45*, 4256. (c) Cízek, J. *Adv. Chem. Phys.* **1969**, *14*, 35. (d) Cízek, J.; Paldus, J. *Int. J. Quantum Chem.* **1971**, *5*, 359. (e) Crawford, T. D.; Schaefer, H. F. *Rev. Comput. Chem.* **2000**, *14*, 33.
5. (a) Roos, B. O.; Taylor, P.; Siegbahn, P. E. *Chem. Phys.* **1980**, *48*, 157. (CASSCF) (b) Ruedenberg, K.; Schmidt, M. W.; Gilbert, M. M.; Elbert, S. T. *Chem. Phys.* 1982, *71*, 41. (FORS)
6. (a)Robb, M. *Mol. Phys.* **1981**, **44**, 173. (b) Langhoff, S. R.; Davidson, E. R. *Int. J. Quantum Chem.* **1974**, *8*, 61.
7. (a) Roos, B. O.; Andersson, K.; Fulscher, M. K.; Malmqvist, P.-A.; Serrano- Andres, L.; Pierloot, K.; Merchan, M. *Adv. Chem. Phys.* **1996**, **93**, 219. (CASPT2) (b) Kozlowski, P. M.; Davidson, E. R. *J. Chem. Phys.* **1994**, *100*, 3672. (MROPT)(c) Hirao, K. *Chem. Phys. Lett.* **1992**, *190*, 374. (d) Hirao, K. *Chem. Phys. Lett.* **1992**, *196*, 397.(e) Hirao, K. *Int. J. Quantum Chem.* **1992**, *S26*, 517. (f). Hirao, K. *Chem. Phys. Lett.* **1993**, *201*, 59. (MRMP2) (g) Nakano, H. *J. Chem. Phys.* **1993**, **99**, 7983.(h) Nakano, H. *Chem. Phys. Lett.* **1993**, *207*, 372. (MCQDPT2)
8. (a) Dunning, T. H; Jr., *J. Chem. Phys*. **1989**, *90*, 1007. (b) S. S. Xantheas; T. H. Dunning, Jr. *J. Phys. Chem.* **1993**, *97*, 18. (c) T. H. Dunning, Jr.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244. (d) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; T. H. Dunning, Jr. *J. Chem. Phys.* **1999** , *110*, 7667. (e) Wilson, A. K.; Woon, D. E.; Peterson, K. A.; T. H. Dunning, Jr. *J. Abstr. Pap.-Am. Chem. Soc*. **1997**, *213*, 60. (f) Kendall, R.

A.; T. H. Dunning, Jr.; Harrison, R. J. *J. Chem. Phys*. **1992**, *96*, 6796. (g) Woon, D. E.; T. H. Dunning, Jr. *J. Chem. Phys*. **1994**, *100*, 2975. (h) Woon, D. E.; T. H. Dunning, Jr. *J. Chem. Phys*. **1995**, *103*, 4572. (i) Peterson, K. A.; T. H. Dunning, Jr. *J. Chem. Phys*. **2002**, *117*, 10548. (j) G3 Basis: Curtiss, L. A.; Redfern, P. C.; Rassolov, V.; Kedziora, G.; Pople, J. A. *J. Chem. Phys*. **2001**, *114*, 9287. (k) Peterson, K.A.; Adler, T.B.; Werner, H.-J. *J. Chem. Phys*. **2008**, *128*, 084102.

9. (a) Pulay, P. *Chem. Phys. Lett*. **100**, *151*, 1983. (b) Saebo S.; Pulay, P. *Chem. Phys. Lett*. **113**, *13*, 1985. (c) Pulay, P.; Saebo, S.*Theor. Chim. Acta* **69**, *357*, 1986. (d) Saebo S.; Pulay, P. *J. Chem. Phys*. **86**, *914*, 1987. (e) Saebo S.; Pulay, P. *Annu. Rev. Phys. Chem*. **44**, *213*, 1993. (f) Saebo, S.; Baker, J.; Wolinski, K.; Pulay, P. *J. Chem. Phys*. **2004**, *120*, 11423. (g) Azhary, A. E.; Rauhut, G.; Pulay, P.; Werner, H. J. *J. Chem. Phys*. **1998**, *108*, 5185. (h) Saebo S.; Pulay, P. *Annu. Rev. Phys. Chem*. **1993**, *44*, 213.

10. (a) Murphy, R. B.; Beachy, M. D.; Friesner, R. A. *J. Chem. Phys*. **1995**, *103*, 1481. (b) Reynolds, G.; Martinez, T.; Carter, E. *J. Chem. Phys*. **1996**, *105*, 6455. (c) Russ N.; Crawford, T. D. *J. Chem. Phys*. **2004**, *121*, 691.

11. (a) Schütz, M.; Hetzer, G.; Werner, H. J. *J. Chem. Phys*. **1999**, *111*, 5691. (b) Werner, H. J.; Manby, F. R.; Knowles, P. J. *J. Chem. Phys*. **2003**, *118*, 8149. (c) Schütz, M.; Hetzer, G.; Werner, H. J. *J. Chem. Phys*. **2001**, *114*, 661. (d) Schütz, M.; Hetzer, G.; Werner, H. J. *Chem. Phys. Lett*. **2000**, *318*, 370. (e) Schütz, M. *J. Chem. Phys*. **2000**, *113*, 9986. (f) Schütz, M. *J. Chem. Phys*. **2002**, *116*, 8772. (h) Schütz, M.; Rauhut, G.; Werner, H. J. *J. Phys. Chem. A* **1998**, *102*, 5997. (i) Rauhut, G.; Azhary, A. E.; Eckert, F.; Schumann, U.; Werner, H. J. *Spectrochim. Acta, Part A* **1999**, *55*, 647. (j) Schütz, M.; Hetzer, G.; Werner, H. J. *J. Chem. Phys*. **1999**, *111*, 5691. (l) Rauhut, G.; Werner, H. J. *Phys. Chem. Chem. Phys*. **2001**, *3*, 4853. (n) Rauhut, G.; Werner, H. J. *Phys. Chem. Chem. Phys*. **2003**, *5*, 2001. (o) Schütz, M.; Hetzer, G.; Werner, H. J.; Lindh, R.; Manby, F. *J. Chem. Phys*. **2004,** *121*, 737. (p) Hrenar, T.; Rauhut, G.; Werner, H. J. *J. Phys. Chem. A* **2006,** *110*, 2060. (q) Hampel, C.; Werner, H. J. *J. Chem. Phys*. **1996,** *104*, 6286.

12. (a) Lee, M.S.; Maslen, P.E.; Head-Gordon, M. *J. Chem. Phys*. **2000,** *112*, 3592.(b) Subotnik, J.E.; Head-Gordon, M. *J. Chem. Phys*. **2005,** *123*, 064108.(b) Subotnik, J.E.; Dutoi, A.D.; Head-Gordon, M. *J. Chem. Phys*. **2005,** *123*, 114108. (c) Subotnik, J.E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys*. **2008,** *128*, 034103. (d) Subotnik, J.E.; Dutoi, A.D.; Head-Gordon, M. *Phys. Chem. Chem. Phys*. **2007,** *9*, 5522. (b) Krylov, A.I.; Sherrill, C.D.; Head-Gordon, M. *J. Chem. Phys*. **2000,** *113*, 6509. (e) DiStasio, R.A.; Jung, Y.; Head-Gordon, M. *J. Chem. Theor. Comput*. **2005**, *1*, 862. (f) Subotnik, J.E.; Head-Gordon, M. *J. Chem. Phys*. **2006,** *125*, 074116.

13. (a) White, C.A.; Head-Gordon, M. *J. Chem. Phys*. **1994,** *101*, 6593. (b) White, C.A.; Johnson, B.G.; Gill, P.M.W.; Head-Gordon, M. *Chem. Phys. Lett*. **1994,** *230*, 8. (c) White, C.A.; Johnson, B.G.; Gill, P.M.W.; Head-Gordon, M. *Chem. Phys. Lett*. **1996,** 248, 482. (d) White, C.A.; Johnson, B.G.; Gill, P.M.W.; Head-Gordon, M. *Chem. Phys. Lett*. **1996,** *253*, 268. (e) White, C.A.; Head-Gordon, M. *Chem. Phys. Lett*. **1996,** *257*, 647. (f) White, C.A.; Head-Gordon, M. *J. Chem. Phys*. **1996,** *105*, 5061. (g) Challacombe, M.; White, C.A.; Head-Gordon, M. *J. Chem. Phys*. **1997,** *107*, 10131.

14. (a) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996,** *271*, 51. (b) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett* **1996,** *248*, 43. (c) Burant, J. C.; Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Chem. Phys. Lett*. **1996,** *258*, 45. (d) Kudin, K. N.; Scuseria, G. E. *Chem. Phys. Lett*. **1998,** *283*, 61. (e) Kudin, K. N.; Scuseria,

G. E. *Chem. Phys. Lett.* **1998,** *289*, 611. (f) Kudin, K. N.; Scuseria, G. E. *J. Chem. Phys.* **1999,** *111*, 2351. (g) Kudin, K. N.; Scuseria, G. E. *Phys. Rev. B* **2000,** *61*, 5141. (h) Kudin, K. N.; Scuseria, G. E. *Phys. Rev. B* **2000,** *61*, 16440. (i) Kudin, K. N.; Scuseria, G. E. J. *Chem. Phys.* **2004,** *121*, 2886.

15. Choi, C.H.; Ruedenberg, K;. Gordon, M.S. *J. Comp. Chem.* **2001,** *22*, 1484.

16. Perez-Jorda, J. M.; Yang, W. *J. Chem. Phys.* **1997,** *107*, 1218.

17. Tobita, M.; Hirata, S.; Bartlett, R. J. *J. Chem. Phys.* **2003,** *118*, 5776.

18. (a) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599. (b) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039. (c) Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. *Chem. Phys. Lett.* **2004**, *394*, 293. (d) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2004,** *120*, 839. (e) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145. (f) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 11386 (g) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903. (h) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Comput. Chem.* **2004**, *25*, 1431. (i) Chen, X.; Zhang, Y.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105. (j) Zhang, D. W.; Zhang, J. Z. H. Int. *J. Quantum Chem.* **2005**, *103*, 246. (k) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2. (l) Li, S.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7251. (m) Jiang, N.; Ma, J.; Jiang, Y. *J. Chem. Phys.* **2006**, *124*, 114112. (n) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103. (o) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *124*, 184703. (p) Mei, Y.; Ji, C.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 094906. (q) Mei, Y.; Wu, E. L.; Han, K. L.; Zhang, J. Z. H. *Int. J. Quantum Chem.* **2006**, *106*, 1267. (r) Mei, Y.; He, X.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins- Struct. Funct. Bioinform.* **2005**, *59*, 489. (s) He, X.; Mei, Y.; Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *Proteins- Struct. Funct. Bioinform.* **2005**, *61*, 423.

19. (a) Imamura, A.; Aoki, Y.; Maekawa, K. *J. Chem. Phys.* **1991**, *95*, 5419. (b) Aoki, Y.; Suhai, S.; Imamura, A. *Int. J. Quantum Chem.* **1994**, *52*, 267. (c) Aoki, Y.; Suhai, S.; Imamura, A. *J. Chem. Phys.* **1994**, *101*, 10808. (d) Mitani, M.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem.* **1995**, *54*, 167; (e) Mitani, M.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem.* **1997**, *64*, 301. (f) Rather, G.; Aoki, Y.; Imamura, A. *Int. J. Quantum Chem.* **1999**, *74*, 35. (g) Gu, F. L.; Aoki, Y.; Imamura, A.; Bishop, D. M.; Kirtman, B. *Mol. Phys.* **2003**, *101*, 1487. (h) Gu, F. L.; Aoki, Y.; Korchowiec, J.; Imamura, A.; Kirtman, B. *J. Chem. Phys.* **2004**, *121*, 10385. (i) Korchowiec, J.; Gu, F. L.; Imamura, A.; Kirtman, B.; Aoki, Y. *Int. J. Quantum Chem.* **2005**, *102*, 785. (j) Korchowiec, J.; Gu, F. L.; Aoki, Y. *Int. J. Quantum Chem.* **2005**, *105*, 875. (k) Makowski, M.; Korchowiec, J.; Gu, F. L.; Aoki, Y. *J. Comput. Chem.* **2006**, *27*, 1603. (l) Orimoto, Y.; Aoki, Y. J. *Polym. Sci. B: Polym. Phys.* **2006**, *44*, 119.

20. (a) Gadre, S. R; Shirsat, R. N.; Limaye, A. C.; *J. Phys. Chem.* **1994**, *98*, 9165. (b) Babu, K.; Gadre, S. R.; *J. Comput. Chem.* **2003**, *24,* 484. (c) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E.; *Theor. Chem. Acc.* **2004**, *111*, 255. (d) Gadre, S. R.; Ganesh, V.; *J. Theor. Comp. Chem.* **2006, 5, 835.**

21. (a) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. *Mol. Phys.* **2005**, *103*, 2255. (b) Kamiya, M.; Hirata, S.; Valiev, M. *J. Chem. Phys.* **2008**, *128*, 074103. (c) Windus,T. L.; Bylaska, E.J.; Dupuis, M.; Hirata, S.; Pollack, L.; Smith, D. M.; Straatsma, T. P.; Aprà, E. In Lecture Notes in Computer Science, vol. 2660, P.M.A. Sloot, D. Abramson, A. Bogdanov, J.J. Dongarra, A. Zomaya, Y. Gorbachev (Eds),

Springer, Berlin (2003). (d) Hirata. S. *J. phys. Chem. A* **2003**, *107*, 9887. (e) Hirata, S. *J. Chem. Phys*. **2004**, *121*, 51. (f) Hirata, S.; Zhan, C.-G.; Aprà, E.; Windus, T.L.; Dixon, D.A. *J. Phys. Chem. A* **2003**, *107*, 10154.

22. (a) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput*. **2007**, *3*, 46. (b) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput*. **20078**, *4*, 1.

23. (a) Tschumper, G. S.; *Chem. Phys. Lett*. **2006**, *427*, 185. (b) Hopkins, B. W.; Tschumper, G. S. *Mol. Phys*. **2005**, *103*, 309. (c) Hopkins, B. W.; Tschumper, G. S. *Chem. Phys. Lett*. **2005**, *407*, 362.

24. (a) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett*. **1999**, *312*, 319. (b) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett*. **1999**, *313*, 701. [(c)] Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett*. **2000**, *318*, 614. (d) Kitaura, K.; Sugiki, S.-I.; Nakano, T.; Komeiji, Y.; Uebayasi, M. *Chem. Phys. Lett*. **2001**, *336*, 163. (e) Inadomi, Y.; Nakano, T.; Kitaura, K.; Nagashima, U. *Chem. Phys. Lett*. **2002**, *364*, 139. (f) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett*. **2002**, *351*, 475. (g) Fedorov, D. G.; Kitaura, K.. *J.Phys. Chem. A* **2007**, *111*, 6904.

25. (a) Deev, V.; Collins, M. A. *J. Chem. Phys*. **2005**, *122*, 154102. (b) Collins, M. A.; Deev, V. *J. Chem. Phys*. **2006**, *125*, 104104. (c) Netzloff, H. M.; Collins; M. A. *J. Chem. Phys*. **2007**, *127*, 134113.

26. (a) Christoffersen, R.; Genson, D. W.; Maggiora, G. M.; *J. Chem. Phys*. **1971**, *54*, 239. (b) Longuet-Higgins, H. C.; Murrel, J. N.; *Proc Phys. Soc*, **1955**, *A68*, 601.

27. (a) Gordon, M.S.; Slipchenko, L.; Hui, L.; Jensen, J.H. *Ann. Rep. Comp. Chem*. **2007**, *3*, 177. (b) Gordon, M.S.; Freitag, M.A.; Bandyopadhyay, P.; Kairys, V.; Jensen, J.H.; Stevens, W.J. *J. Phys. Chem*. **2001**, *105*, 293.

28. Day, P.N.; Jensen, J.H.; Gordon, M.S.; Webb, S.P.; Stevens, W.J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys*. **1996**, *105*, 1968.

29. Adamovic, I.; Gordon, M.S. *J. Phys. Chem*. **2005**, *109*, 1629.

30. Webb, S.P.; Gordon, M.S. *J. Phys. Chem*. A **1999**, *103*, 1265.

31. Yoo, S.; Gordon, M. S. *J. Chem. Phys*. **2008**, *129*, 14112.

32. (a) Stone, A. J. Alderton, M. *Mol. Phys*. **1985**, *56*, 1047. (b) Stone, A.J., The Theory of Intermolecular Forces; Oxford Press; **1996**.

33. (a) Edmiston, C.; Ruedenberg, K. Rev. Mod. Phys. **1963**, *35*, 457. (b) Raffenetti, R. C.; Ruedenberg, K.; Janssen, C. L.; Schaefer, H. F. III. *Theor. Chim. Acta* **1993**, *86*, 149.

34. (a) Jensen, J.H; Gordon, M.S. *Mol. Phys*. **1996**, *89*, 1313. (b) Murrell, J. N. *Proc. Roy. Soc. (Lond*.) **1965**, *A284*, 566. (c) Murrell, J. N. *J. Chem. Phys*. **1967**, *47*, 4916.

35. (a) Adamovic, I.; Gordon, M.S. *Mol. Phys*. **2005**, *103*, 379, (b) Amos, R.D.; Handy, N.C.; Knowles, P.J.; Rice, J.E.; Stone, A.J. *J. Phys. Chem*. **1985**, *89*, 2186. (c) Piecuch, P. *Mol. Phys*. **1986**, *59*, 1085.

36. (a) Li, H.; Gordon, M.S.; Jensen, J.H. *J. Chem. Phys*. **2006**, *124*, 214107. (b) Li, H.; Gordon, M.S. *Theor. Chem. Accts*. **2006**,*115*, 385.

37. (a) Slipchenko, L.; Gordon, M.S. *J. Comp. Chem*. **2006**, *28*, 276. (b) Freitag, M.A.; Gordon, M.S.; Jensen, J.H.; Stevens W.J. *J. Chem. Phys*. **2000**, *112*, 7300.

38. Tang, K.T.; Toennies, J.P. *J. Chem. Phys*. **1984**, *80*, 3726.

39. Slipchenko, L.; Gordon, M. S. *submitted*. (b) Murrell, J. N.; Teixeira-Dias, J. J. C. *Mol. Phys*. **1970**, *19*, 521.

40. (a) Hehre, W.J.; Ditchfield, R.; Pople, J.A. *J. Chem. Phys*. **1972**, *56*, 2257 (b) Francl, M.M.; Pietro, W.J.; Hehre, W.J.; Binkley, J.S.; Gordon, M.S.; DeFrees, D.J.; Pople, J.A. *J.Chem.Phys*. **1982**, *77*, 3654. (c) Hariharan, P.C.; Pople, J.A. *Theoret.Chim.Acta* **1973**, *28*, 213.

41. (a) Hariharan, P. C.; Pople, J. A. *Theoret.Chim.Acta* **1973**, *28*, 213. (b) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J.Chem. Phys*. **1980**, *72*, 650. (c) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. *J.Comp.Chem*. **1983**, *4*, 294.

42. (a) 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.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem*. **1993**, *14*, 1347. (b) Gordon, M. S.; Schmidt, M.W., Theory and Applications of Computational Chemistry, the first forty years; Elsevier; Amsterdam; **2005**.

43. Allen, F.H. *Acta Cryst*. *B* **2002**, *58*, 380.

44. (a) Boys, S.F.; Bernardi, F. *Mol. Phys*. **1970**, *19*, 553. (b) Ostlund, N.S.; Merrifield, D.L. *Chem. Phys. Lett*. **1976**, *39*, 612.

45. Schwenke, D. W.; Truhlar, D. G. *J. Chem. Phys*. **1985**, *82*, 2418.

46. Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev*. **1994**, *94*, 1887.

Table 1. In (a), all non-bonded interactions (kcal/mol) are calculated using the indicated level of theory. In (b), CCSD(T) is used in the EFP column for all non-bonded interactions with separations ≤ 2.7 Å

	MP2	CCSD(T)	EFP[†]	EFP[§]
(a) non-bonded	-3.38	-4.14	-11.95	-5.96
(b) non-bonded	--	--	-3.30	-3.77

† Tang-Tonnies damping of the dispersion term
§ overlap damping of the dispersion term

Table 2. Timings for calculating the 129 interactions in ZOVGAS. In the column labeled Mix, CCSD(T) was employed for all interactions with nearest atom- atom distances ≤ 2.7 Å.

	CCSD(T)	EFP	Mix
Hours	5038	4.1	~500

Table 3. Relative MP2/6-311++G(3df,2p) energies (kcal/mol) of retinal cis-trans isomers with fully *ab inito* (MP2) and SFM level 3 using non-bonded level 3 (SFM). MP2/6-311++G(3df,2p) was employed for all interactions with nearest atom- atom distances ≤ 2.7 Å.

	MP2	SFM
trans Retinal	0.0	0.0
11-cis Retinal	3.9	4.1

Table 4. Absolute SFM HF/6-31++G(d,p) errors in isomerization energies (kcal/mol) for alpha helixes (see Figure 4) relative to *ab initio*. Fragmentation level 3 is used with non-bonded level 2 (NB2) and Level 3 (NB3). HF/6-31++G(d,p) was employed for all interactions with nearest atom- atom distances $\leq$ 2.7 Å. Fully HF SFM non-bonded absolute errors ($d_{min} = \infty$) in parentheses. Fully EFP absolute errors ($d_{min} = 0$) in square brackets.

	NB2	NB3
	HF/6-31++G(d,p)	HF/6-31++G(d,p)
Isomer	HF kcal/mol	HF kcal/mol
MAQWUW_1-MAQWUW_2	2.5 (1.1) [2.5]	2.2 (0.2) [1.9]
WUYCUO-WUYDAV	2.6 (5.8) [1.7]	2.1 (6.1) [10.7]
WUYCUO-WUYDEX	3.4 (2.5) [2.9]	0.3 (3.1) [4.1]
YETPES_1-YETPES_2	1.9 (5.3) [4.1]	0.3 (1.0) [3.1]
Mean Absolute Error	2.6 (3.7) [2.8]	1.2 (2.6) [4.9]
Mean Absolute Standard Error	0.6 (3.7) [1.0]	1.0 (2.6) [3.9]

Figure 1. Pictorial examples of level 1 fragmentation for ethanol and ethanal. The first step breaks bonds creating functional groups; hydrogens are capped at a chemically reasonable bond distance in the direction of the broken bond.
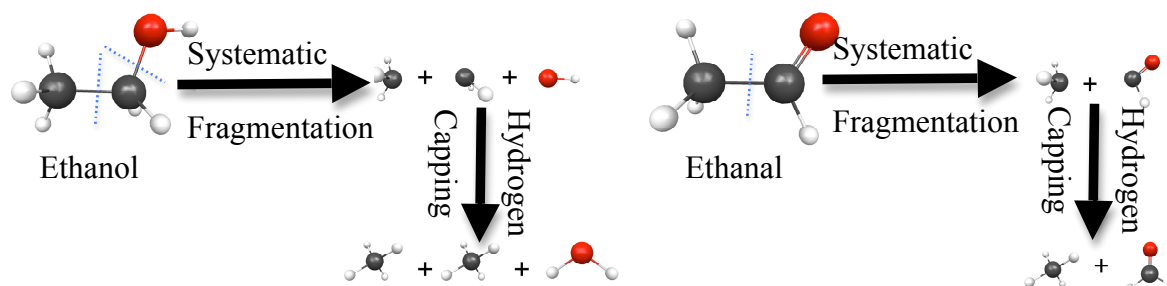
Figure 2. The ZOVGAS molecule used as a test for EFP non-bonded interactions. Structure is from the Cambridge Structural Database (CSD)
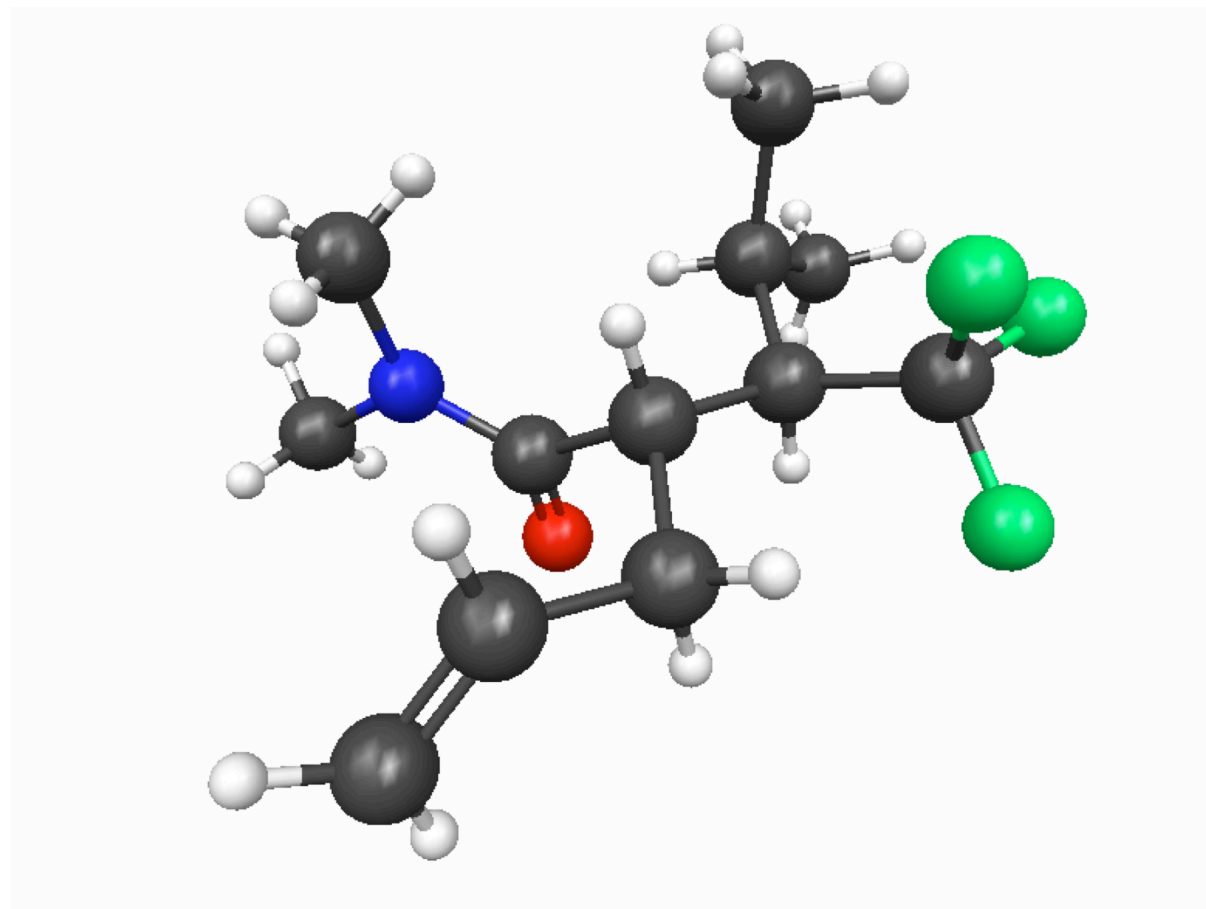
Figure 3. Depiction of retinal isomers used in Table 3. Structures are from the Cambridge
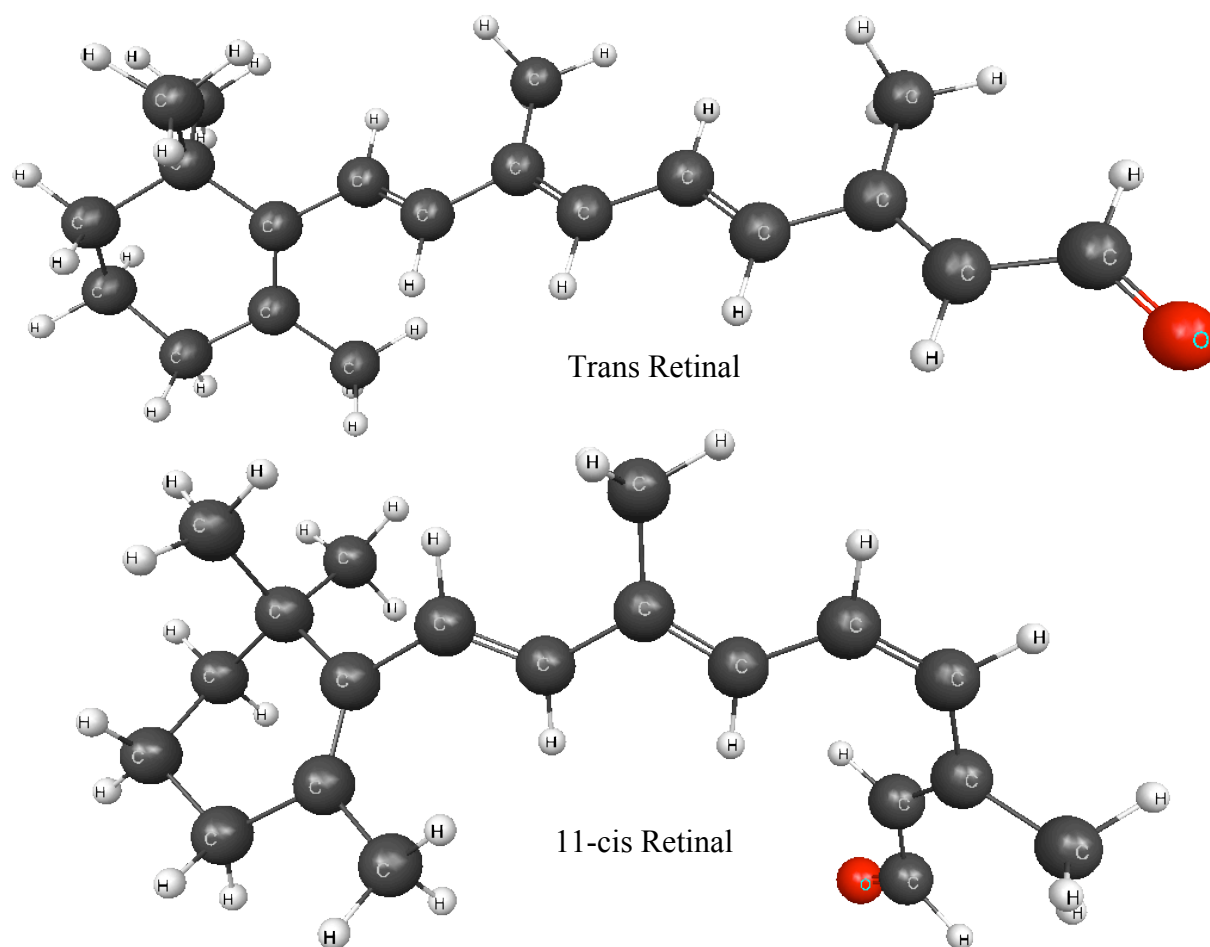Structural Database (CSD).
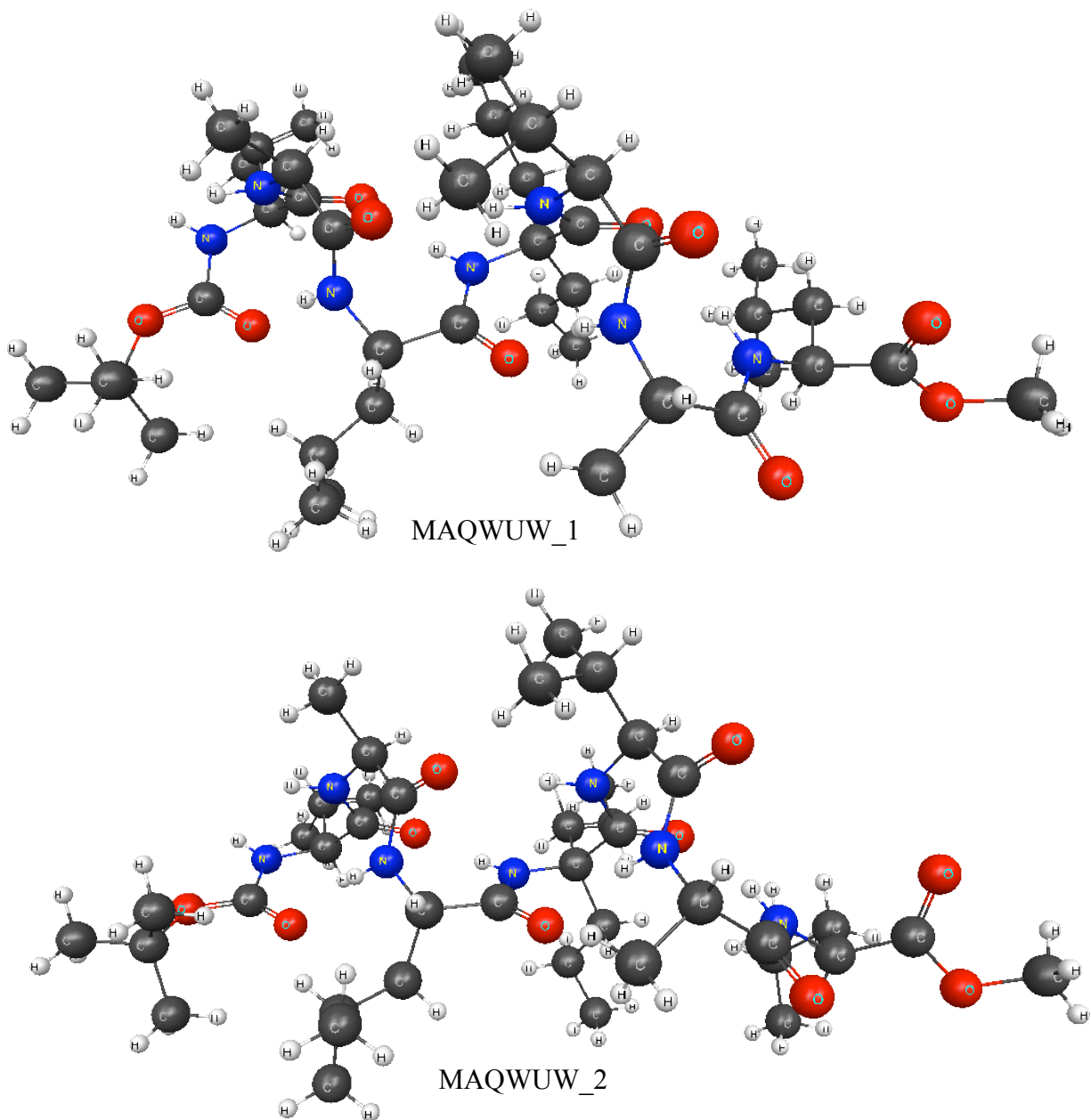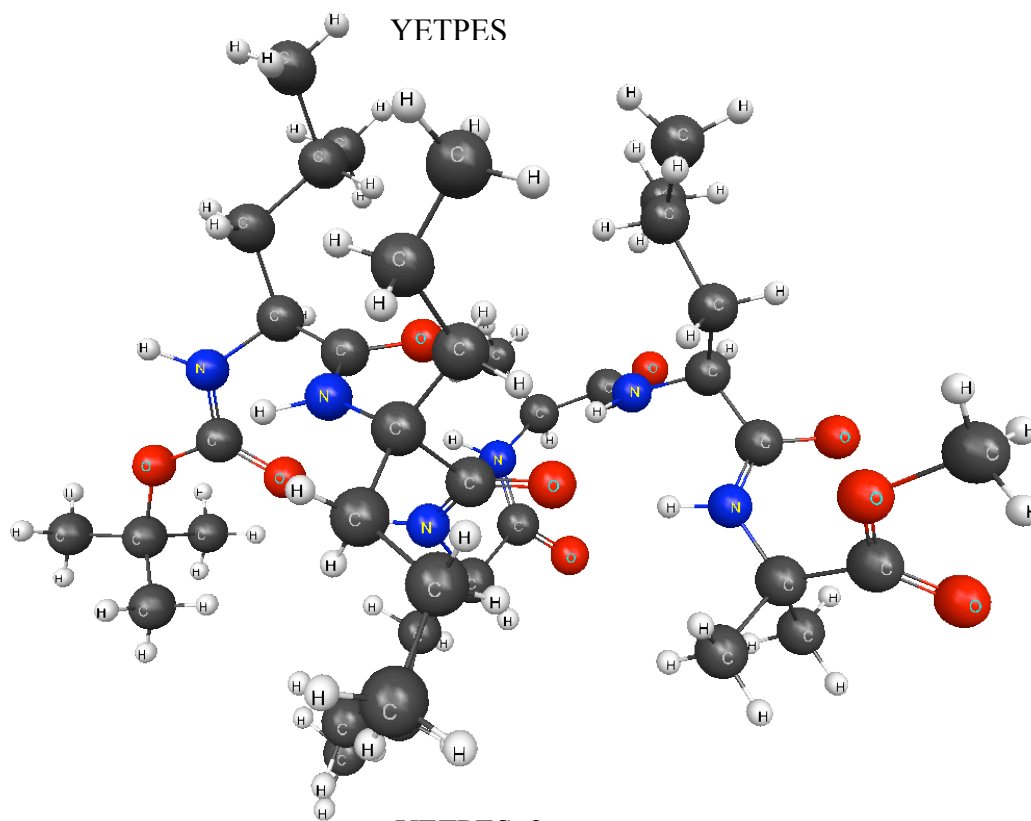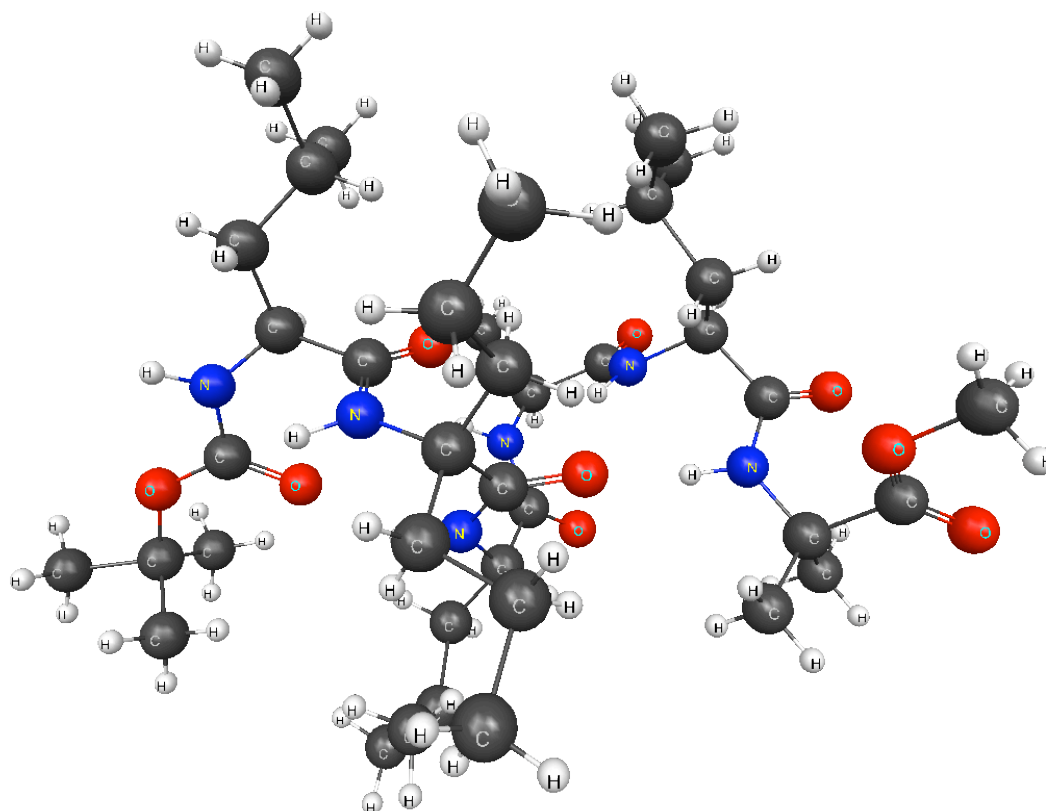

Trans Retinal

11-cis Retinal

Figure 4. Depiction of alpha helix isomers used in Table 4. Structures are from the Cambridge Structural Database (CSD).

MAQWUW_1

MAQWUW_2

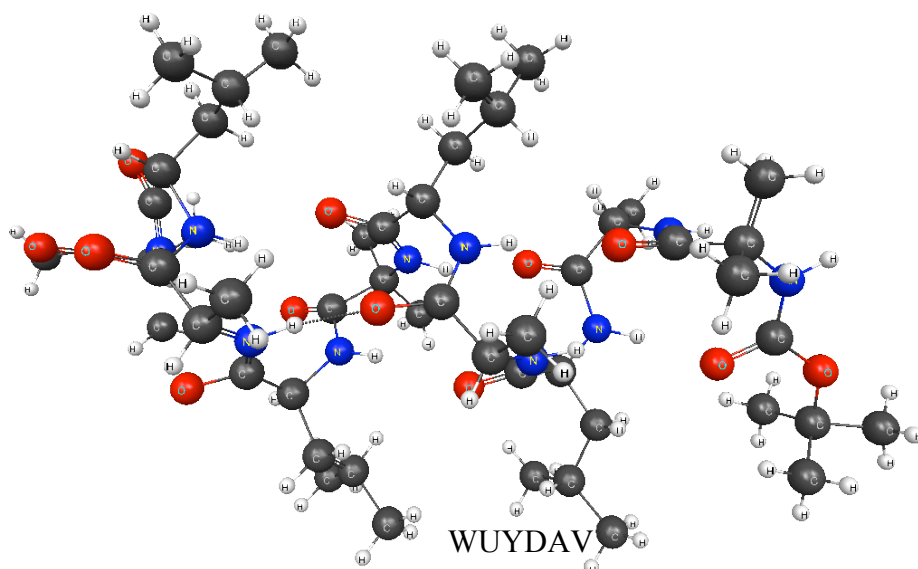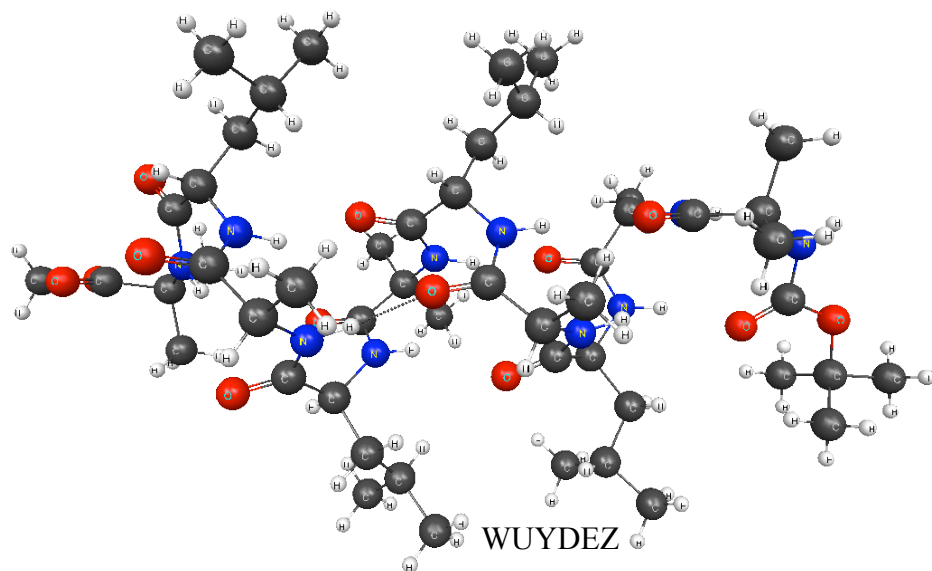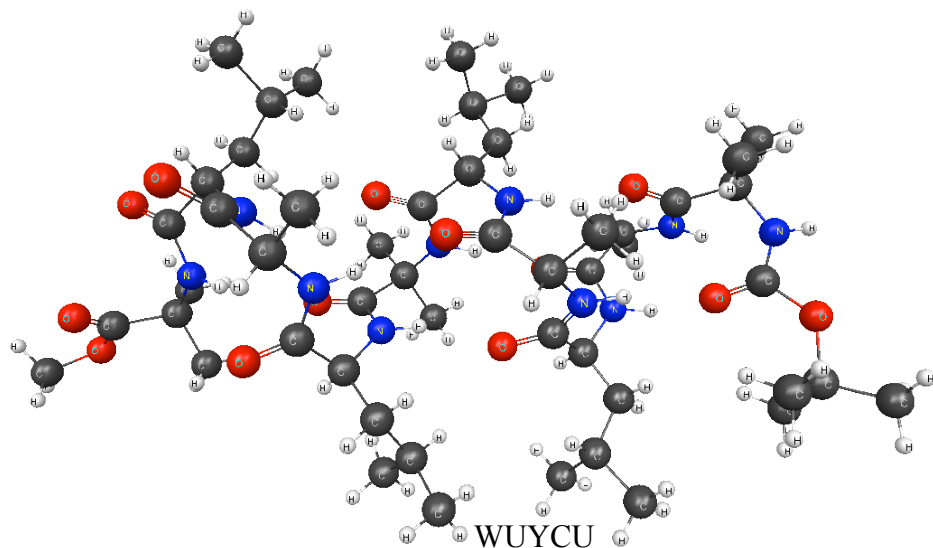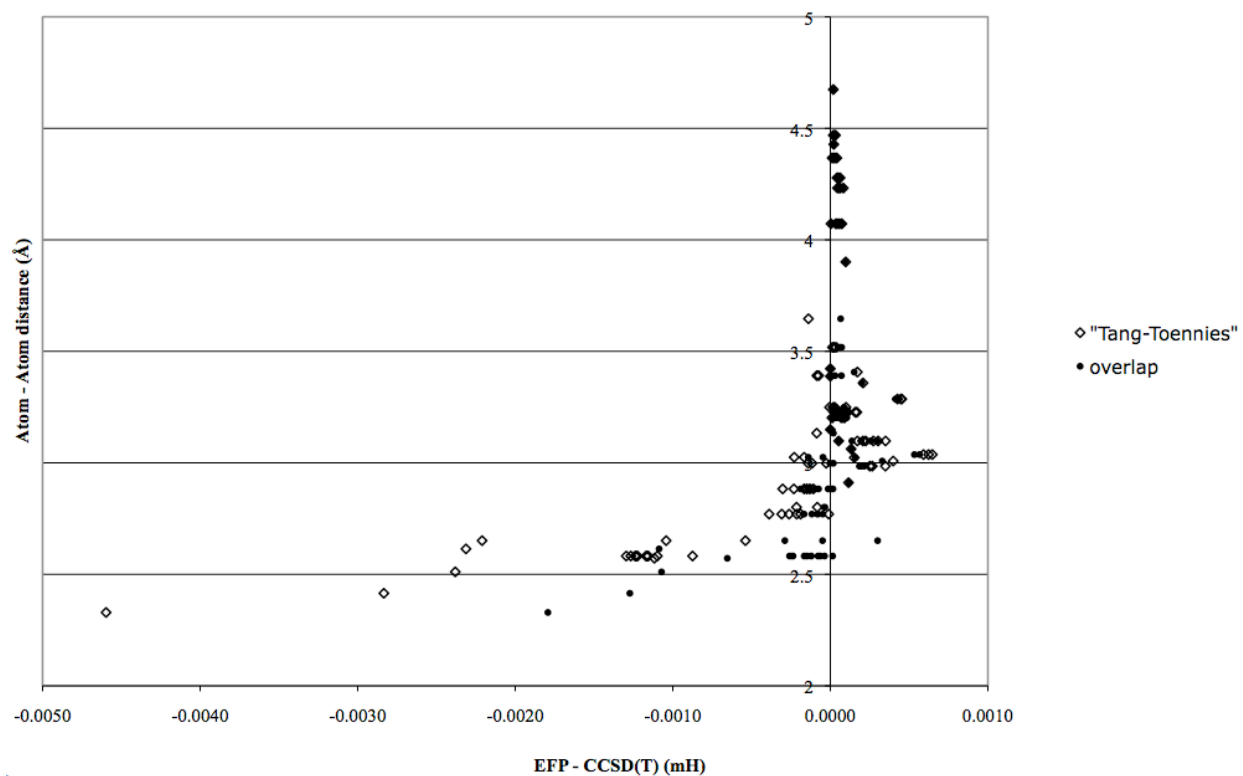YETPES

YETPES_2

WUYCU

WUYDEZ

WUYDAV

Figure 5. A plot of the 129 non-bonded level 3 interactions of ZOVGAS. The nearest atom-atom distance (Å) is plotted vs. the EFP interaction energy minus the CCSD(T) interaction energy. Both the Tang-Tonnies (◊)dispersion damping scheme and the overlap (•) dispersion damping scheme are shown.

MOLECULAR DYNAMICS SIMULATIONS WITH THE EFFECTIVE FRAGMENT POTENTIAL METHOD

Jonathan M. Mullin and Mark S. Gordon*

Department of Chemistry, Iowa State University, Ames, Iowa 50011

*mark@si.msg.chem.iastate.edu

Abstract

1. Introduction

Condensed phase phenomena have increasingly become the focus of computational chemistry given that environmental effects can be important to chemical and biological processes. Methods that can scale from clusters to the condensed phase without limiting accuracy are necessary for reliable prediction of bulk properties. Atomistic molecular dynamics (MD) simulations require accurate solvent models that can accurately predict relevant bulk properties such as density, local structure, compressibility, diffusion, and dielectric constant. Discrete solvent models are usually parameterized to reproduce bulk properties[1] or due to computational cost are more appropriate for clusters than for large MD simulations.

The effective fragment potential (EFP)[2] method was originally designed and implemented to describe discrete solvent effects. The initial focus was on the accurate prediction of the effect of solvents on chemical reactions and on the study of small clusters of water molecules. The EFP method has been shown to accurately reproduce cluster properties,[3] solvent mediated reactions,[4] global minima obtained from Monte Carlo simulations,[5] and preliminary reports illustrated that EFP/MD can reproduce radial distribution functions of water, benzene, and CCl₄ reasonably well.[6] The present work

introduces the theory and evaluation of several properties via EFP/MD simulations using water and argon as examples.

2. Theoretical and Computational Details

2.1 EFP (Effective Fragment Potential)

The effective fragment potential (EFP) has two implementations, (EFP1) a water-specific model with empirically fitted parameters, and (EFP2) a general model for any molecular system that has no empirically fitted parameters. The EFP1 model has been implemented at the HF,[7] DFT,[8] and MP2[9] levels of theory. EFP1 has been integrated with HF,[7] DFT,[8] MCSCF,[10] singly excited configuration interaction (CIS),[11] and time-dependent density functional theory (TDDFT).[12] The EFP1 method contains one-electron potentials, which may be added to an *ab initio* electronic Hamiltonian of a solute. There are three interaction energy terms in EFP1, corresponding to Columbic (electrostatic) interactions, induction (polarization), and a third term fitted to the water dimer potential. This last term contains those interactions that are not accounted for in the first two: exchange repulsion + charge transfer. In the MP2 version of EFP1, dispersion is treated as an additional term. The Coulombic and induction terms are determined from monomer QM calculations. The EFP1 formulation for an EFP solvent molecule μ and a QM coordinate s as follows:

$$V_{el}(\mu,s) = \sum_{k=1}^{K} V_k^{Elec}(\mu,s) + \sum_{l}^{L} V_l^{Pol}(\mu,s) + \sum_{m}^{M} V_m^{\text{Re}\,m}(\mu,s)$$

$$(1)$$

The three terms on the right hand side of Eq. (1) represent the Coulomb, induction, and remainder terms, respectively, where K, L and M are the number of corresponding expansion points.

The Coulomb interaction is obtained via the Stone distributed multipolar analysis.[13] This expansion is truncated at the octopole term. In EFP1 K=5 expansion points for the water molecule (nuclear centers and bond midpoints).

The polarization/induction term in Eq. (1) describes the interaction of an induced dipole on one fragment (or solute molecule) with the permanent dipole on another fragment (or solute molecule), expressed in terms of the dipole polarizability. Truncating at the first (dipole) term in the polarizability expansion is viable, since the molecular polarizability tensor is expressed as a tensor sum of localized molecular orbital[14] (LMO) polarizabilities. For water, four such LMOs are used: two O lone pairs, and two O-H bonds. The polarization energy is then iterated to self-consistency (within the SCF cycles if a quantum solute is present), so it is able to capture some many body effects

The remainder term includes exchange repulsion, charge transfer, and some short-range correlation contribution (in EFP1/DFT). The dispersion (in EFP1/MP2) is a separate fourth term. These interactions are represented by a linear combination of Gaussian functions expanded at the atom centers for EFP-QM interactions. The EFP-EFP interactions are expanded at atom centers and the center of mass via a single exponential to capture the angular dependence of the charge transfer contribution.[7] The coefficients and exponents of the Gaussian and exponential functions were optimized, by fitting to

water dimer structures, chosen to represent a selection of water-water orientations and O-O distances.[7]

In EFP2[2a] all terms are derived from first principles with no empirically fitted parameters, therefore it can be used to represent any molecule of interest. The interaction energy includes Coulomb, polarization, exchange-repulsion,[15] dispersion[16] and charge transfer.[17]

$$E = E_{coul} + E_{ind} + E_{exrep} + E_{disp} + E_{ct}. \tag{2}$$

The first two of these terms are evaluated in exactly the same manner as in EFP1. The five terms in the EFP2 potential may be grouped into long range, $(1/R)^n$ distance dependent, or short range interactions, which decay exponentially. The Coulombic, induction, and dispersion are long-range interactions, whereas the exchange repulsion and charge transfer are short range.

The exchange repulsion interaction between two fragments is derived as an expansion in the intermolecular overlap.[15] When this overlap expansion is expressed in terms of frozen LMOs on each fragment, the expansion can reliably be truncated at the quadratic term. Since, once the EFP has been generated, the basis set is used only to calculate overlap integrals as the inter-fragment geometry changes, the computation is very fast and quite large basis sets are realistic.

Dispersion interactions are often expressed by an inverse R expansion,

$$E_{disp} = \sum_n C_n R^{-n} \tag{3}$$

where the coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities integrated over the entire frequency range.[16] The first term in the expansion, n=6, corresponds to the induced dipole-induced dipole (van der Waals) interactions. In the EFP2 method, this term is evaluated using the time-dependent HF method. In addition the contribution of the n=8 term is estimated.[16a] The C_6 coefficients are derived in terms of interactions between pairs of LMOs, one each on the two interacting fragments.

The charge transfer interaction is derived using a supermolecule approach, in which the occupied valence molecular orbitals on one fragment are allowed to interact with the virtual orbitals on another fragment. This interaction term leads to significant energy lowering in *ab initio* calculations on ionic or highly polar species when incomplete basis sets are employed.[17]

Treating the Coulomb, induction and dispersion terms with a classical approximation does not correctly model short-range interactions where quantum mechanical charge densities begin to overlap. Therefore, each term is multiplied by a damping (screening) expression. These short distances cause the classical Coulombic interactions do become too repulsive, while the induction is too attractive. The unphysical behavior is avoided by augmenting the electrostatic multipoles with exponential damping functions of the form:

$$f_{damp} = 1 - \exp(-\alpha R) \tag{4}$$

where parameters α are determined at each multipole expansion point by fitting the multipole damped potential to reproduce the Hartree–Fock potential.[18] For induction,

both exponential damping, as in Eq. (2) and Gaussian damping are effective, but the Gaussian damping seems to be more generally applicable and is therefore recommended. Dispersion interactions formally decrease to zero at short range; therefore, the Tang-Toennies damping function Eq. (5) is used: [19]

$$f_n(R) = 1 - e^{(-bR)} \sum_{k=0}^{n} \frac{(bR)^k}{k!} \tag{5}$$

In the EFP model n is set to 6, and b is set to the same value for all LMO-LMO pairs[19]. A new approach based on the overlap integrals between interacting fragments[20] appears to be promising and will be used in subsequent applications. In future EFP applications, the overlap-based dispersion damping is recommended, because it appears to be more generally effective.

It is useful to consider the relative costs of the five EFP interaction terms. Based on relatively small molecules and taking the cost of the Coulomb and dispersion terms to be one unit, the induction interaction costs approximately two units, exchange repulsion costs about five units, and charge transfer costs ~10 units. For larger molecules, the relative costs of exchange repulsion and charge transfer will decrease since they will scale linearly in the large molecule limit. The EFP method is basis set dependent, since a basis set is used to generate the multipoles and the molecular polarizability tensor, and to evaluate the exchange repulsion and charge transfer interactions.

The EFP model is currently a rigid body model potential, so the internal geometries of fragments are fixed. Within this constraint, analytic gradients for all terms have been derived and implemented, so full intermolecular geometry optimizations can

be performed. Because the method involves no empirically fitted parameters, an EFP for any system can be generated by a "makefp" run in the GAMESS[21] suite of programs.

2 .2 MD simulations

The simulations, unless otherwise stated, were performed using 64 molecules with periodic boundary conditions in the NVT ensemble. The temperature (T) was maintained at 298.16 K for water and 85.0 K for argon, by employing a Nosé-Hoover thermostat.[22] The simulation density was chosen to be 0.997 g/cm^3 for water and 1.375 g/cm^3 for Ar. The equations of motion were integrated with the velocity Verlet algorithm[23] with a time step of 1 fs over 500ps. The first 100 ps were treated as an equilibration. The molecule-molecule interaction energies were treated with a simple switching function.[6b] The switching function shown in Eq. (6) shifts the pairwise potential function smoothly to zero in the range of 5 - 6.2 Å for water-water interactions and 6.0 - 9.0Å for Ar-Ar interactions.

$$if \quad r_{c.m.} \leq r_a, S = 1$$

$$if \quad r_a < r_{c.m.} < r_b, S = 1 - 10\left(\frac{r_{c.m.}^2 - r_a^2}{r_b^2 - r_a^2}\right)^3 + 15\left(\frac{r_{c.m.}^2 - r_a^2}{r_b^2 - r_a^2}\right)^4 - 6\left(\frac{r_{c.m.}^2 - r_a^2}{r_b^2 - r_a^2}\right)^5$$

$$if \quad r_b < r_{c.m.}, S = 0 \tag{6}$$

In Eq. (6) $r_{c.m.}$ is the distance between the centers of masses of two interacting molecules. r_a and r_b are the starting and ending points, respectively, of the switching function. When $r_{c.m.} = r_a$ and $r_{c.m.} = r_b$, both the first and second derivatives of S are zero.

Cutoffs for the Coulombic interactions (charge-charge, charge-dipole, dipole-dipole, and charge-quadrupole) were treated via Ewald summation.[24] The EFPs were created with the 6-31++G(d,p) basis set.[25] All calculations were done with the

GAMESS[21] suite of programs. The Ar dispersion uses Tang-Toennies damping (Eq. (5)) with b=2.8. For water b=3.0. The overlap-damped dispersion is reported for $g_{OO}(r)$ only. Induction damping has been used in all the simulations. The geometry used for the water models has an OH bond distance of 0.948636 Å and an HOH bond angle of 106.0732˙.

2.3 Bulk Property evaluation

2.3.1 Diffusion

Diffusion may be calculated in two ways.[26] First, one can use the slope of the time dependent mean-squared displacement from the initial coordinates. This is represented by the Einstein equation:

$$2tD = \frac{1}{3}\left\langle |r_i(t) - r_i(0)|^2 \right\rangle, \tag{7}$$

r(t) is the center of mass position vector at time t, r(0) are the initial x,y,z center of mass coordinates, D is the diffusion constant. Alternately the velocity auto correlation function may be used to calculate the diffusion constant.

$$D = \frac{1}{3}\int_0^\infty dt \left\langle V_i(t) \bullet V_i(0) \right\rangle, \tag{8}$$

$V_i(t)$ is the velocity of particle i at time t, $V_i(0)$ is the initial x,y,z center of mass velocity of particle i, and D is the diffusion constant. The values reported and discussed below use the velocity auto correlation function.

2.3.2 Heat of Vaporization

The heat of vaporization is obtained using the equation:[27]

$$\Delta H_{vap} \approx -\langle E \rangle + RT \ , \tag{9}$$

E is the energy of the liquid system, R is the gas constant and T is the simulation temperature.

2.3.3 *Static Dielectric Constant*

Using the total dipole moment of the system, M, one can calculate the static dielectric constant ε_0 via:

$$\varepsilon_0 = 1 + \frac{4\pi}{3\bar{V}k_B T}(\langle M^2 \rangle - \langle M \rangle^2) \ , \tag{10}$$

k_B is the Boltzmann constant, V is the average volume of the system, and T is the simulation temperature. Convergence of dielectric constants is slow often taking several ns.[28]

2.3.4 *Coordination number*[29]

Using a radial distribution function the coordination number can be calculated via:

$$n_c = \frac{N}{V} 4\pi \int\limits_0^{R_c} r^2 g(r) dr \ , \tag{11}$$

N is the number of water molecules in the system; V is the volume within a radial cutoff of R_c. $g(r)$ is the radial distribution function. The coordination number, n_c, is the number of molecules coordinated with a central molecule.[29] For water $R_c = 3.3$Å was used for the $g_{OO}(r)$, while for argon, $R_c = 5.5$ Å. These R_c were chosen based on previously published work.[33,43]

2.3.5 *Kirkwood factors (infinite and finite)[30]*

The Kirkwood factors are temperature dependent dielectric properties. The finite Kirkwood factor is a fluctuation formula using the equilibrium squared dipole moment fluctuations:

$$G_k = \frac{\langle M^2 \rangle - \langle M \rangle^2}{N \langle \mu^2 \rangle}, \tag{12}$$

μ is the atomic dipole moment, M is the system dipole moment. The infinite Kirkwood factor depends directly on the static dielectric constant (ε_0) and the finite Kirkwood factor (G_k) via:

$$g_k = \frac{2\varepsilon_0 + 1}{3\varepsilon_0} G_k. \tag{13}$$

2.3.6 *Isothermal compressibility[31]*

Finite differencing is used to calculate the isothermal compressibility via:

$$\kappa_T = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_T \approx -\left(\frac{\ln(\rho_2/\rho_1)}{T_2 - T_1}\right)_T, \tag{14}$$

$\rho_1 \approx \rho_0 + 0.04$ and $\rho_2 \approx \rho_0 - 0.04$. ρ_0 is the equilibrium density (0.997 g/cm^3) at 298.15K and 1 atm. This required two NVT simulations, run for 100ps. The value reported is an average over the previous 80ps.

2.3.7 *Hydrogen bond lifetime*

Using the Luzar-Chandler model[32] for the hydrogen bond lifetime, the following correlation functions were calculated.

$$c(t) = \frac{\langle h(0)h(t) \rangle}{\langle h \rangle} \tag{15}$$

$$n(t) = \frac{\langle h(0)[1-h(t)]H(t) \rangle}{\langle h \rangle} \tag{16}$$

h(t) denotes if a tagged pair of molecules is hydrogen bonded (h(t)=1) or not (h(t)=0) at time t. The criterion of a hydrogen bond is defined as an $O-H\cdots O$ angle of greater than 150° and O-O distance less than 3.5Å. h(0) uses these definitions for initial conditions. $\langle h \rangle$ is the averaged value of h(t) over the simulation. H(t)=1 if the oxygen-oxygen distance of the tagged molecules is less than a given cutoff (3.5Å) at time t, otherwise H(t)=0. For long times, $-\frac{dc}{dt} = kc(t) - k'n(t)$ and the hydrogen bond lifetime is given by $\tau_{HB} = 1/k$. Time derivatives of c can be calculated from the simulation by Eq. (15),

$$\frac{dc}{dt} = \frac{\langle h(0)[1-h(t)] \rangle}{\langle h \rangle} \quad . \tag{17}$$

Hydrogen bond lifetime calculation requires an NVE simulation, which was run for 100ps. The average temperature over the simulation was 300k.

3. Results and Discussion

3.1 *Water*

Figure 1 compares the water radial distribution functions $g_{OO}(r)$ for EFP1/MP2 and for EFP2 with several variants of basis sets and dispersion damping. The MP2 version of EFP1 was chosen because it includes dispersion and therefore provides a more accurate $g_{OO}(r)$ than EFP1/HF and EFP1/DFT.[6a] In cluster studies larger basis sets have been shown to improve the ability of EFP2 to reproduce the results of ab initio calculations.[18a] However, preliminary molecular dynamics (MD) simulations suggest no significant basis set effects, so modest basis sets are employed here. Recently, a dispersion damping approach that is based on the intermolecular (inter-EFP) overlap was implemented for the EFP2 method.[20] The results obtained using both the Tang-Toennies and overlap damping of dispersion are displayed in Figure 1.

The blue curve in Figure 5 uses Tang-Toennies dispersion damping (b=3.0 in Eq. (5)) with the 6-31++G(d,p) basis set. The overlap dispersion damping is examined using both the 6-31++G(d,p) (red curve) and 6-31G(d,p) (orange curve) basis sets. This comparison illustrates the importance of including diffuse functions in the basis set, in order to obtain an accurate $g_{OO}(r)$. If overlap damping and the 6-31++G(d,p) basis set is used, EFP2 nearly reproduces the X-ray $g_{OO}(r)$.[20c] If diffuse functions are omitted the two *peaks in the calculated $g_{OO}(r)$ are in the correct positions, but the first peak is too high and the intermediate well is too deep. The EFP1/MP2 $g_{OO}(r)$ is in reasonably good agreement with the experimental curve. This is satisfying, but as noted earlier, EFP1 is only applicable to water, while EFP2, with no empirically fitted parameters, is applicable to any system and is therefore much more general. The fixed geometry can influence

property evaluation. In the case of the static dielectric constant, structure of a model can have a large effect.[33]

The predictions of EFP2 bulk properties, obtained from the MD simulations, are based on the Tang-Toennies dispersion damping. These results, as well as those obtained using EFP1/MP2 and the simpler SPC/E[1a] potential, are compared with the experimental values in Table 1. All of the computational methods are in reasonable agreement with experiment. Each of the three methods – SPC/E, EFP1/MP2 and EFP2 are in closer agreement with experiment for some properties, but none of the three methods stands out as superior to the other two. Both EFP models were able to nearly reproduce the bulk dipole moment.[34] The predicted coordination numbers[33] are in good agreement with experiment as are the $g_{OO}(r)$.[35] Both EFP models overestimate the heat of vaporization compared to expirement;[36] however, the EFP heat of vaporization agrees well with the recent work of Fanourgakis, Schenter, and Xantheas.[27] They extrapolated the quantum effects of water based on small clusters to a centroid MD simulation. This may suggest that the semi-classical nature of the EFP model potential can capture some of the quantum effects lacking in classical model potentials. Diffusion constants of 2.42×10^{-5} $cm^2 s^{-1}$ and 2.17×10^{-5} $cm^2 s^{-1}$ was found for EFP1/MP2 and EFP2 compared to an experimental value of 2.3×10^{-5} $cm^2 s^{-1}$.[37] The relative permittivities of 72 and 85 predicted by EFP1/MP2 and EFP2 are in good agreement with the experimental value of 78.5.[38] The Kirkwood factors predicted by both EFP methods compare well with both experiment and the SPC/E values.[38] The isothermal compressibility predicted by both EFP1/MP2 (4.46×10^{-5} atm^{-1}) and EFP2 (4.84×10^{-5} atm^{-1}) are in reasonable agreement with the experimental value of 4.84×10^{-5} atm^{-1}.[39] The average hydrogen bond lifetime,

apparently not known experimentally for water, is predicted by EFP1/MP2, EFP2, and SPC/E to be 2.1ps, 2.5ps, and 2.3ps, respectively.[33] The performance of EFP2 in predicting these properties is particularly encouraging, given the generality of this method.

3.2 *Argon*

The Tang-Toennies (b=2.8) dispersion damping facilitates good accuracy of the liquid argon system; simulations with the overlap dispersion damping are currently being run. The radial distribution function for $g_{Ar\text{-}Ar}(r)$ is shown in Figure 3 for EFP2 and compared with the experimental neutron scattering data.[40] Theory and experiment are in reasonably good agreement, with the EFP2 peak and valley slightly shifted to larger values of r. Comparisons to bulk properties are shown in Table 2. Again, there is reasonable agreement between EFP2 and experiment, with EFP2 slightly over-estimating the heat of vaporization and the diffusion constant. The EFP2 coordination number is predicted to be 8.8 compared to 8.8-10.4 for several Ar models and 11.4 at the argon triple point for R_c=5.5 Å.[42]

4. Conclusion

The effective fragment potential is able to reproduce bulk properties of liquid water for both EFP1/MP2 and the general EFP2 model as accurately as SPC/E, which was been fitted to reproduce bulk properties. The inclusion of diffuse functions in the EFP2 model was necessary to nearly reproduce the $g_{OO}(r)$ when using the overlap dispersion damping. Initial tests appear to show that other properties in MD will be similar to those presented here for the overlap damping. The EFP2 method predicts slightly larger heats of vaporization and diffusion coefficients compared to experiments for liquid argon. The

EFP method has now successfully predicted bulk behavior in liquid systems in addition to previous use as an accurate approach for predicting explicit solvent effects on small clusters.

5. Acknowledgements

References.

1. Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381. (spc) (b) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269. (spc/e ?)
2. Gordon, M.S.; Slipchenko, L.; Hui, L.; Jensen, J.H. *Ann. Rep. Comp. Chem.* **2007**, *3*, 177. (b) Gordon, M.S.; Freitag, M.A.; Bandyopadhyay, P.; Kairys, V.; Jensen, J.H.; Stevens, W.J. *J. Phys. Chem.* **2001**, *105*, 293.
3. (a)Aikens, C. M.; Gordon, M. S. *J. Am. Chem. Soc.* **2006**, *128*, 12835. (glycine) (b) Mullin, J. M.; Gordon, M. S. *J. Am. Chem. Soc.* **2009**, *submitted*. (alanine) (c) Wood, G.P.F.; Gordon, M.S.; Radom, L.; Smith, D.M. *J. Phys. Chem.* **2009**, *submitted*. (d) Bandyopadhyay, P.; Gordon, M.S. J. Chem. Phys. **2000**, *113*, 1104. (e) Jensen, J.H.; Gordon, M.S. *J. Am. Chem. Soc.* **1995**, *117*, 8159.
4. (a) Webb, S. P.; Gordon, M. S. *J. Phys. Chem. A* **1999**, *103*, 1265. (menshuken) (b) Adamovic, I.; Gordon, M. S. *J. Phys. Chem.* **2005**, *109*, 1629. (SN2_EFP1DFT)
5. Day, P.N.; Pachter, R.; Gordon, M.S.; Merrill, G.N. *J. Chem. Phys.* **2000**, *112*, 2063. (b) Metropolis, N.; Rosenbluth, A.; Teller, A. *J. Chem. Phys.* **1953**, *21*, 1087.
6. Netzloff, H.M.; Gordon, M.S. *J. Chem. Phys.* **2004**,*121*, 2711. (b) Li, H.; Netzloff, H.M.; Gordon, M.S. *J. Chem. Phys.* **2006**, *125*, 194103. (c) Pranami, G.; Slipchenko, L.V.; Lamm, N.H.; Gordon, M.S. *Multi-scale Quantum Models for Biocatalysis: Modern Techniques and Applications.* edited by T.-S. Lee and D. M. York Springer; Verlag; **2009**.
7. Day, P.N.; Jensen, J.H.; Gordon, M.S.; Webb, S.P.; Stevens, W.J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.
8. Adamovic, I.; Gordon, M.S. *J. Phys. Chem.* **2005**, *109*, 1629.
9. Song, J.; Gordon. M. S. unpublished.
10. Webb, S.P.; Gordon, M.S. *J. Phys. Chem. A* **1999**, *103*, 1265.
11. Pooja, A.; Gordon, M. S.; in preparation

12. Yoo, S.; Zahariev, F.; Sok, S.; Gordon, M. S.; *J. Chem. Phys*. **2008**, *129*, 14112.

13. Stone, A. J.; Alderton, M. *Mol. Phys*. **1985**, *56*, 1047. (b) Stone, A. J. *The Theory of Intermolecular Forces.* Oxford University Press, London, UK, **1996.**

14. (a) Edmiston C.; Ruedenberg, K. *Rev. Mod. Phys*. **1963**, *35*, 457. (b) Raffenetti, R. C.; Ruedenberg, K.; Janssen, C. L.; Schaefer, H. F. III. *Theor. Chim. Acta* **1993**, *86*, 149.

15. (a) Jensen, J.H; Gordon, M.S. *Mol. Phys*. **1996**, *89*, 1313. (b) Murrell, J. N. *Proc. Roy. Soc. (Lond.)* **1965**, *A284*, 566. (c) Murrell J. N. *J. chem. Phys.* **1967**, *47,* 4916.

16. (a) Adamovic, I.; Gordon, M.S. *Mol. Phys*. **2005**, *103*, 379. (b) Amos, R.D.; Handy, N.C.; Knowles, P.J.; Rice, J.E.; Stone, A.J. *J. Phys. Chem*. **1985**, *89*, 2186. (c) Piecuch, P. *Mol. Phys*. **1986,** 59, 1085.

17. Li, H.; Gordon, M.S., *Theor. Chem. Accts*. **2006**, *115*, 385. (b) Li, H.; Gordon, M.S.; Jensen, J.H. *J. Chem. Phys*. **2006**, *124*, 214107.

18. (a) Slipchenko, L.; Gordon, M.S. *J. Comp. Chem*. **2006**, *28*, 276. (b) Freitag, M.A.; Gordon, M.S.; Jensen, J.H.; Stevens W.J. *J. Chem. Phys*. **2000**, *112*, 7300.

19. Tang, K.T.; Toennies, J.P. *J. Chem. Phys*. **1984**, *80*, 3726.

20. (a) Slipchenko, L.; Gordon, M. S. in preparation. (b) Murrell, J. N.; Teixeira-Dias, J. J. C. *Mol. Phys*. **1970**, *19*, 521.(c) In the final form of this paper the overlap dispersion damping will be used.

21. (a) 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.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., Jr. *J. Comput. Chem*. **1993**, *14*, 1347. (b) Gordon, M.S.; Schmidt, M.W., Theory and Applications of Computational Chemistry, the first forty years; Elsevier; Amsterdam; 2005

22. Nosé, S. *Mol. Phys*. 1984, **52**, 255. B) Hoover, W. G. *Phys. Rev. A* **1985**. *31*, 1695. (NooseHoover)

23. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys*. **1982**, *76*, 637. (velocity verlet)

24. (a) Berthaut, F. *J. de Physique Rad*. **1952**, *13*, 499-505. (b) Slipchenko, L.V.; Gordon. M.S. *J.Comput.Chem*. **2007**, *28*, 276. (Ewald Sums)

25. (a) Hariharan, P. C.; Pople, J. A. *Theoret. Chim. Acta* **1973**, *28*, 213. (b) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys*. **1980**, *72*, 650. (c) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. *J. Comp. Chem*. **1983**, *4*, 294.

26. Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids* Oxford Science, Oxford. **1987**. (Diffusion)

27. Fanourgakis, G. S.; Schenter, G. K.; Xantheas. S. S. *J. Chem. Phys.* **2006**, *125*, 141102. (Heat of VAP)

28. Neumann, M. *Mol. Phys*. **1983**, *50*, 841. (b) Jepsen, D. W. *J. Chem. Phys. Lett*. **2000**, *258*, 121. (Dielectric constant)

29. Soper, A. K. *Chem. Phys. Lett.* **2000**, *258*, 121. (Coord #)

30. Neumann, M. *Mol. Phys*. **1986**, *57*, 97. (Kirkwood factors)

31. Motakabbir, K. A.; Berkowitz, M. *J. Phys. Chem*. **1990**, *94*, 8359. (Isothermal compressibility)

32. (a) Luzar, A.; Chandler, D. *Nature* **1996**, *379*, 55. (b) Luzar, A.; Chandler, D. *Phys. Rev. Lett*. **1996**, *17*, 928. (c) Xu, H.; Stern, H. A.; Berne, B. J. *J. Phys. Chem. B* **2002**, *166*, 2054. (H-bond life)

33. Yujie, W.; Tepper, H. L.; Voth, G. A. *J. Chem. Phys*. **2006**, *124*, 024503. (coord #)

34. Badyal, Y. S.; Saboungi, M.-L.; Price, D. L.; Shastri, S. D.; Haeffner, D. R.; Soper, A. K. *J. Chem. Phys*. **2000**, *112*, 9206. (bulk dipole)

35. Sorenson, J. M. *et al. J. Chem. Phys*. **2000**, *113*, 9149. (expirmental g(oo)

36. Wagner, W.; Pruss, A. *J. Phys. Chem*. **2002**, *31*, 387. (heat of vap)

37. Bieze, T. W. N.; van der Maarel, J. R. C.; Leyte, J. C. *Chem. Phys. Lett.* **1993**, *216*, 56. (diffusion, alpha)

38. *Handbook of Chemistry and Physics*, edited by R. C. Weast CRC, Cleveland, **1977**. (heat capactiy) (static dielectric) (kirkwood)

39. VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*, 014515. *(*isothermal compressability)

40. Yamell, J.L.; Katz, M.J.; Wenzel, R.G.; Koenig, S.H. *Phys. Rev. A* **1973**, *7*. 2130. *Ar radial distribution*

41. Naghizadeh, J.; Rice, S. A. *J. Chem. Phys*. **1962**, *36*, 2710. *Ar diffusion*

42. Brostow,W.; Sicotte, Y. *Physica A,* **1975***, 80*, 513. *Ar Coordination number*

43. Sychev, V. V.; Vasserman, A. A.; Kozlov, A. D.; Spiridonov, G. A.; Tsymarny, V. A., *Thermodynamic Properties of Neon, Argon, Krypton, and Xenon* New York, Hemisphere, **1987**. (Ar Heat of Vap)
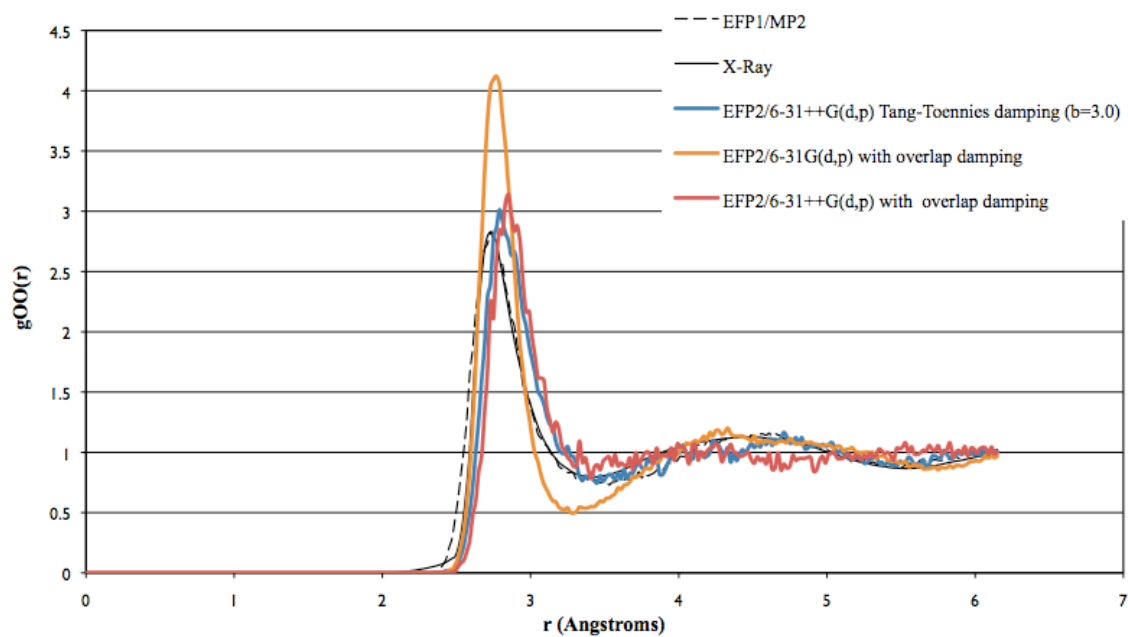
Figure 1. The $g_{OO}(r)$ radial distribution function for 64 water molecules.

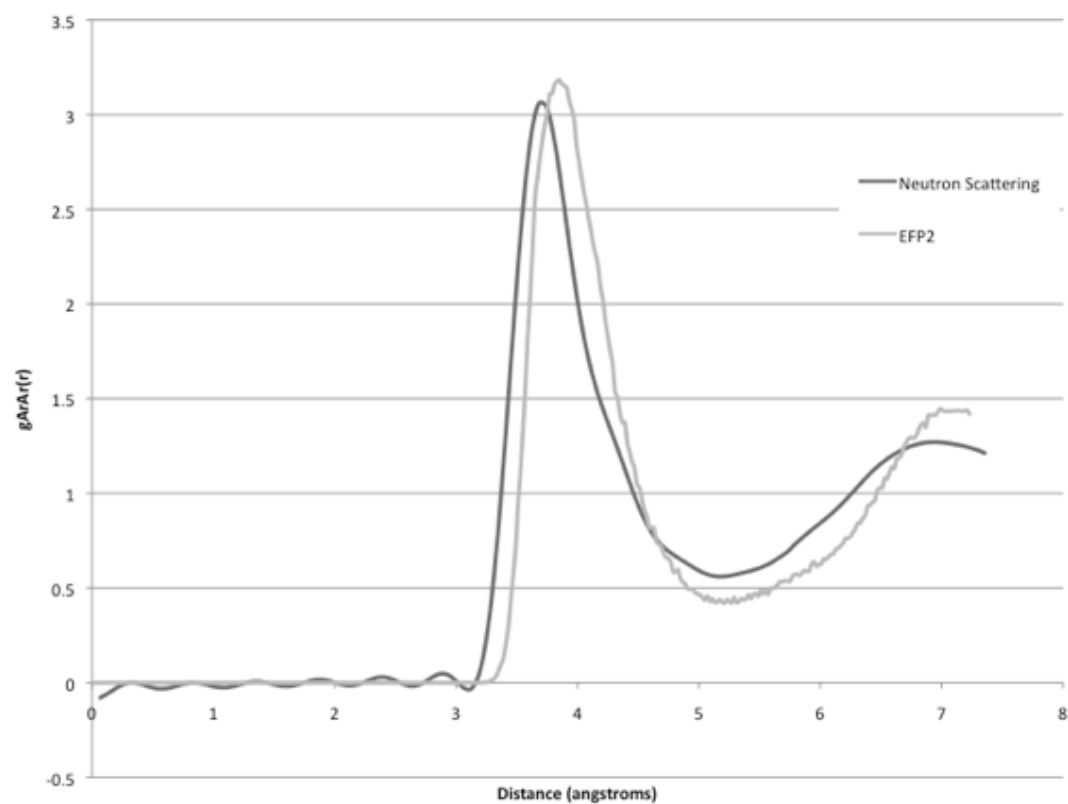Figure 2. Radial distribution function $g_{ArAr}(r)$ for 64 argon molecules represented by the EFP2 method.

Table 1. Bulk properties of water, EFP/MP2 and EFP2 (Tang-Toennies damping, b=3.0) models compared to experiment and SPC/E. Properties and EFP models are defined in the computation details.[a]

	units	EFP1/MP2	EFP2	SPC/E[b]	Exp.
$<\mu>$	D	2.8	3.1	2.41	2.9[g]
n_c	$R_c=3.3$ Å	4.32	4.45	4.34	4.26[b]
ΔH_{vap}	kcal/mol•K	11.2	11.61	10.76	10.52[c]
D_s	10^{-5} cm^2 s^{-1}	2.42	2.17	2.41	2.3[d]
e_o		72	85	76.66	78.5[e]
G_k		4.84	4.16	4.02	--
g_k		3.25	2.79	2.70	2.9[e]
κ_t	10^{-5} atm^{-1}	4.46	4.84	4.46	4.58[f]
τ_{HB}	ps	2.1	2.5	2.3	--

[a]In the final form of this paper the overlap dispersion damping will be used.

b Ref 33

c Ref 36

d Ref 37

e Ref 38

f Ref 39

g Ref 34

Table 2. Bulk properties of argon, EFP2 model and experiment. Properties and EFP2 model are defined in the computation details.

	units	EFP2	Exp.
n_c	R_c=5.5 Å	8.8	8.8[a]
ΔH_{vap}	kcal/mol•K	1.53	1.38[b]
D_s	10^{-5} cm^2 s^{-1}	2.07	1.84[c]

a Ref 42
b Ref 43
c Ref 41

GENERAL CONCLUSIONS

Early theoretical chemistry research was focused on obtaining very accurate gas phase phenomena. Building on that foundation of systematically improvable *ab inito* methods and basis sets, forays into condensed phase phenomena have been the concentration of much research in the area for some time. Herein, the chapters of this dissertation have shown a breadth and depth of study within condensed phase theoretical chemistry.

 In chapters three and four, the systematic solvation of alanine has allowed the determination of the liquid phase enthalpy of zwitterion formation from neutral alanine. Furthermore the effects on the water structures during the solvation process were explored. It is through such a systematic study that convergence of these properties may be evaluated. Often, when solvating a species, an ill-defined concept such as a first solvation shell is employed. Here precisely where the solvation shell becomes the dominant structure was presented along with an ensemble of representative structures through the addition of 1 -49 waters. The number of waters required to solvate the neutral form of alanine is in the range of 46-49 waters, while the zwitterion exhibits solvation at the addition of 42 waters.

Chapters five and six focus not on QM/MM methods, but rather the use of fragmentation schemes which achieve near linear scaling with ~1kcal/mol accuracy relative to *ab initio* methods. While QM/MM methods have expanded the size of systems that are accessible to computations, the use of classical model potentials for the description of the environment can be a limiting factor, given that the electron density of the MM region and its impact on the QM region is not usually properly accounted for. Both of the fragmentation schemes employed, FMO and SFM, allow the investigation of bonded systems, and the clustered model of SFM remains a work in progress. By employing linear scaling methods, larges

systems such as proteins, surfaces, and clusters can be explored via *ab initio* methods while retaining accuracy of ~1kcal/mol relative to the full *ab initio* calculation. Further development of fragmentation schemes within a molecular dynamics system would allow linear scaling elucidation of condensed phase chemical phenomena.

In chapter seven the use of molecular dynamics(MD) as a means to capture time dependent phenomena of liquid systems is presented. Several properties such as pressure, diffusion, dielectric constant, and free energy are now available within the GAMESS suite. Several more properties, such as heat capacity and thermal expansion can be calculated via finite differencing by the end user. The EFP methods used within MD calculations have been shown to accurately capture the dynamical nature of liquid water; furthermore the EFP2 method has been shown to describe argon and should be applicable to the investigation of other liquid systems.

The solvation of alanine has answered many of the basic questions. However, extension of this work to other amino acids may be considered. The transition from N to Z forms should not vary greatly (for uncharged R groups) from the requirement of seven waters since the majority of waters form along the hydrogen bonding NH_3^+ and COO^- groups. Where there is a charge center on the R group more waters may be required. It was shown that in order to determine the enthalpy of the N to Z alanine discretely solvated species of both forms is required. Investigation into this enthalpy for other amino acids does not need to be systematically, only to complete a first solvation shell with a continuum model as a third layer.

The extension of the molecular dynamics properties can be extended to include the contributions from molecules treated via quantum mechanics. These extensions should not be

overly complicated, but would most likely require further work on the linear scaling methods such as FMO and SFM in order to generate a simulation with statistical significance. Temperature of each substituent should be available not just the entire ensemble in order to avoid spurious temperature distributions. Visualization and calculation of external properties can be included in MacMolPlt or via programs such vmd which support GAMESS.

Generalized free energy calculations such and alchemical perturbation or potential of mean force may be added to the MD code. A main concern thus far in using umbrella sampling and force biasing has been the self-consistent polarization term. It may not be necessary to have the DIPIT subroutine term kill a MD run. This term may not converge in the process of umbrella sampling unless great care is taken. Alchemical perturbation can be implemented by extended the current ability to turn on and off individual terms of the EFP2 calculation. By using control over each EFP the interaction of an EFP can slowly be turned on and off, so long as the center of mass coordinates of the two fragments used in the alchemical perturbation are both are the same. Alchemical perturbation is preferable for use where the perturbation is in a constrained environment, while potential of mean force is preferable for use in systems with high degrees of freedom.

The SFM method can be extended to include gradient and hessians via the same process of simple summation used for the energy. Extension of this methodology to clusters has proven difficult in identifying the source of deviation from *ab inito* calculations. Systems with large polarization effects may not be effectively treated by the current EFP implementation for the non-bonded energy. A first order response may be needed to fully capture these interactions. The SFM method may be used as a model in order to define

fragments automatically for other fragmentation methods such as the fragmentation

molecular orbital method.

Biographical Sketch

Jonathan Mullin was born the 23rd of November 1981 in Council Bluffs, Iowa. Jonathan currently is a graduate student in the field of theoretical chemistry. After completing a B.S (2004) and M.S.(2005) degree in biochemistry, Jonathan entered the Iowa State university chemistry program under the direction of Professor Mark Gordon. Mr. Mullin's research interests include solvation chemistry, biological systems, and QM/MM methods.

Jonathan M. Mullin
201 Spedding, Iowa State University, Ames, Iowa 50011

Education

Iowa State University, Ames, Iowa.   08/05-present
> Physical Chemistry, Doctoral Studies

Iowa State University, Ames, Iowa.   05/04-08/05
> Masters of Science in Biochemistry

Iowa State University, Ames, Iowa.   08/00-08/04
> Bachelor of Science in Biochemistry

Research Experience:

Graduate Research Assistant, Department of Chemistry   08/05-present
> Iowa State University and Ames Laboratory
> Research Advisor: Prof. Mark S. Gordon
> Research Area: Computational Quantum Chemistry

Graduate Research Assistant, Department of Biochemistry   05/04-08/05
> Iowa State University and Ames Laboratory
> Research Advisor: Prof. Mark S. Gordon, Richard Honzatko
> Research Area: Computational Quantum Chemistry

Undergraduate Research Assistant, Department of Chemistry   01/01-05/04
> Iowa State University and Ames Laboratory
> Research Advisor: Prof. Mark S. Gordon
> Research Area: Computational Quantum Chemistry

Research Assistant, Department of Zoology and Genetics   05/98-08/99
> Iowa State University
> Research Advisor: Prof. Eric Henderson
> Research Area: Protozoan Transformational Genetics

Recent Research Interests:

- The microsolvation of amino acids (alanine) with water (1-1024) , along with how the addition of water changes the favorability of neutral (gas phase) to zwitterion (in solution).
- The electronic structure interactions of protein substrate binding, are examined, and this information is used to develop novel binding substrates for proteins of unknown structure.
- Development of new environmentally friendly high energy fuels by investigation of polynitrogen species using dynamic reaction path methods.
- Expansion of the molecular dynamics (MD) code in GAMESS to include a complete set of property calculations and ensembles, along with completing an interface for QM/MM MD calculations.
- Study of the behavior of amino acids on the silicon (100) surface using high level hybrid quantum mechanics/molecular mechanics methods. The potential energy surfaces for reactions between amino acids are described.
- Development of fast hybrid theoretical methods towards a goal of treating biological systems.

Professional Experience:

Visiting Scholar, Australian National University   09/08
EFP:Theory and Practical notes. (Workshop) Sydney University, Australia   03/07
Visiting Scholar, Australian National University   01/07-05/07
Preparing Future Faculty   '06
Teaching Assistant, General and Physical Chemistry, Iowa State University   08/04-05/05

Tutor for Organic and Physical Chemistry, Calculus based Physics   08/02-05/04

Papers:

Vibrational Spectroscopy for Glycine adsorbed on Silicon Clusters: Harmonic and Anharmonic Calculations for Models of the Si(100)-2x1 Surface. D. Shemesh, J. Mullin, M. Gordon, R. B. Gerber, Chemical Physics, 2008, 347, 218-228.

Accurate methods for large molecular systems. Mark S. Gordon, Jonathan M. Mullin, Spencer R. Pruitt, Luke B. Roskop, Lyudmila V. Slipchenko, and Jerry A. Boatz, *Submitted*. J. Phys. Chem.

Alanine:Then there was Water. J. Mullin, and M. Gordon, *Submitted*. J. Phys. Chem.

Alanine:From puddle to ponds. J. Mullin, and M. Gordon, *in preperation*.

Systematic Molecular Fragmentation and the Effective Fragment Potential: An Efficient Method for Capturing Molecular Energies. J. Mullin, M. Collins, L. Roskop, M. Gordon, *in preperation*.

Molecular Dynamics with Generalized Effective Fragment Potentials. J. Mullin, M. Gordon, *in preperation*.

Free Energy methods with the Effective Fragment Potential. J. Mullin, M. Gordon, *in preperation*.

On the Energetics of polynitrogen Species. J. Mullin, and M. Gordon, *in preperation*.

Presentations:

Molecular Dynamics with Generalized Effective Fragment Potentials. J. Mullin, M. Gordon, Poster, WATOC, Sydney, NSW, AU

Systematic Fragmentation of Clustered Systems. J. Mullin, M. Gordon, Poster, American Conference of Theoretical Chemistry(ACTC), Northwestern University

Molecular Dynamics with Generalized Effective Fragment Potentials. J. Mullin, M. Gordon, Talk, 235th ACS National Meeting, New Orleans, LA

What if all the kings men COULD put Humpty Dumpty back together Again? J. Mullin,M. Collins, M. Gordon, Poster, 235th ACS National Meeting, New Orleans, LA

Hybrid Methods for Accurate Electronic Structure Calculations. J. Mullin,M. Collins, M. Gordon, Poster, 234th ACS National Meeting, Boston, MA

Alanine:Then there was Water. J. Mullin, and M. Gordon, Talk, Royal Australian Chemical Institute Organic/Physical Conference, Adelaide, Australia

Alanine:Then there was Water J. Mullin, and M. Gordon, Poster session, Practicing Chemistry with theoretical tools, Maui, HI

Alanine:Then there was Water. J. Mullin, and M. Gordon, Talk, 232nd National ACS meeting, San Francisco, CA

PolyNitrogen Species:Fuel or Fizzle. J. Mullin, and M. Gordon, Talk, 232nd National ACS meeting, San Francisco, CA

Si(100):When Glycine Attacks. J. Mullin, and M. Gordon, Physical Chemistry poster session & SciMix 232nd National ACS meeting, San Francisco, CA

PolyNitrogen Species:Fuel or Fizzle. J. Mullin and M. Gordon, Talk, 39th Midwest Theoretical Chemistry Conference, Columbus, OH

poster presentations (posters previously shown), Iowa State University Chemistry Department Open House (02/04, 02/05, 02/06, 02/08)

Alanine: Then there was Water. J. Mullin and M. Gordon, Talk, 38th Midwest Theoretical Chemistry Conference, Columbia, MO

Alanine:Then there was Water. J. Mullin and M. Gordon, Physical Chemistry poster session 230th National ACS meeting, Washington, DC

Competitive substrates of transmembrane proteins as a means to design new substances. J. Mullin, R. Martin, and M. Gordon, Physical Chemistry poster session & SciMix 228th

National ACS meeting, Philadelphia PA

Characterization of Alanine in Neutral and Zwitterion forms. J. Mullin, M.S. Gordon, poster presentation, REU poster presentations, Ames, IA.

Continuum Solvation Models and Amino Acids J.. Mullin, and M. S. Gordon, Argonne undergraduate symposium.

Tertrahymena Transformation, new approaches. J. Mullin, E Henderson , STTG symposium Iowa State University

<u>Honors and Awards:</u>

Nelson Fellowship	08/08-07/09
GAANN Fellowship	08/07-07/08
Lindau meeting of Nobel Laureates, young researcher participant	06/06
Frank J. Moore and Thoreen Beth Moore Fellowship	05/06
Chemistry Department Academic Scholarship, Iowa State University	08/04-05/07
Vocational Rehabilitation Scholarship, Iowa State University	08/00-05/09
ACS Travel Awards	'05 '07 '08
Chevron Phillips Travel Award	'05 '08
Proctor and Gamble Travel Award	'06

<u>Professional Societies:</u>

American Chemical Society
American Physical Society