
אור ותמים
LUX ET VERITAS

Abstract

Continuum Solvation Models and Force Field Development
for Computer-Aided Drug Design

John Philip Terhorst

2011

A description of our implementation of the generalized Born / surface area (GB/SA) solvation model with free-energy perturbation (FEP), including an approximation used in calculating the total Born energy of the system, is presented. Our approximation is based on the assumption that a significant number of pairwise energy calculations may be omitted with little-to-no impact on the total change in energy of the system after a Monte Carlo move because the impact of a moving atom on the Born radius of a distant atom is small. Thus, we structured our implementation of GB/SA in such a way that the Born energy between an unmoving pair of atoms is only recalculated after a move if the Born radius of either atom has changed by more than a specified threshold since the last accepted move. Prior benchmarks demonstrated that existing GB/SA methodologies were insufficient for the purposes of calculating free energies of binding, and FEP simulations with GB/SA solvation were too computationally expensive to be used with any practicality. With our approximation, improved efficiency was achieved while affording minimal error: the influence of our approximation on accuracy of free energies of binding was negligible, with any error introduced by the approximation falling well below the statistical error of the Metropolis Monte Carlo algorithm, and speed-up of up to 62% was observed. The conclusion is that with our approximation, GB/SA is a viable solvent choice for FEP of large systems. Comparison between GB/SA and TIP4P in a substituent scan was quantitative to qualitative, with free energies of binding usually in agreement within 1 kcal/mol, producing the same substitution pattern on a drug candidate found to give high anti-retroviral activity as predicted by previous simulations with TIP4P explicit water.

In Chapter 2, thermochemical data obtained from G3B3 quantum mechanical calculations are presented for 18 prototypical organic molecules that exhibit E/Z conformational equilibria. The results are fundamentally important for molecular design including the evaluation of structures from protein–ligand docking. For the 18 E/Z pairs, relative energies, enthalpies, free energies, and dipole moments are reported; the $E - Z$ free-energy differences at 298 K range from +8.2 kcal/mol for 1,3-dimethyl carbamate to −6.4 kcal/mol for acetone oxime. A combination of steric and electronic effects can rationalize the variations. Free energies of hydration were also estimated using the GB/SA continuum solvent model. These results indicate that differential hydration is unlikely to qualitatively change the preferred direction of the E/Z equilibria.

In Chapter 3, torsion parameters for the OPLS-AA force field were developed and fit to quantum mechanical data for 65 prototypical derivatives of benzene, pyridine, furan, thiophene, and pyrrole, containing methyl, ethyl, isopropyl, cyclopropyl, *t*-butyl, vinyl, hydroxy, methoxy, thio, methylthio, amino, *N*-methylamino, and *N*,*N*-dimethylamino substituents. The parameterization yielded well-defined torsion curves that mimicked the quantum mechanically calculated curves to less than 0.5 kcal/mol error and afforded 66.9%–95.8% reduction of error over unparameterized molecular mechanical curves. For a small subset of the 65 molecules (OH, SH, and SCH_3 derivatives of 2-pyridine and OH and OCH_3 derivatives of 2-furan), a shift in dihedral angle population was observed in the transfer from gas-phase to aqueous environments owing to significantly different torsion profiles between gas and GB/SA. Results contribute to the development of molecular modeling software and an understanding of small-molecule conformational energetics and dynamics: well-predicted dihedral populations for a given molecular species can aid computational, medicinal, and organic chemists in making the correct choices in molecular designs to achieve the desired molecular shape in a given environment. The rationale for the choice of heterocycles and derivatives in the context of drug design was given in Chapter 1 and the need for such development was highlighted in Chapter 2.

Continuum Solvation Models and Force Field Development for Computer-Aided Drug Design

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
John Philip Terhorst

Dissertation Director: William L. Jorgensen

December, 2011

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I want to thank my advisor, Professor William (Bill) Jorgensen, for the opportunity to pursue a Ph.D. under his guidance. It has been an honor and a privilege to work under his direction for these many years and for this experience I feel amazingly fortunate. He has taught me, both consciously and unconsciously, the many virtues of being a good scientist. His unwavering dedication and high standards in the oftentimes frustrating pursuit of some of the field's most daunting challenges are indeed to be both admired and imitated. I also appreciate his contribution of time, ideas, and funding to make my Ph.D. both successful and rewarding.

The members of the Jorgensen lab have contributed immensely to my professional and scientific growth during my time at Yale. Particular gratitude is expressed to Dr. Julian Tirado-Rives for his almost-infinite patience, valuable advice, and valued friendship. It is rare to encounter an individual with the depth and breadth of knowledge that Julian has and I can say with absolute certainty that my experience in graduate school would not have been what it was without his help and support.

Additional thanks goes to Paty Morales de Tirado for her ever-positive attitude and never-ending encouragement. Her support and care helped me to overcome setbacks and to stay focused on my work. Her contribution of Juanito to the lab also gets an honorable mention.

Dr. Julien Michel's willingness to share his expertise in Monte Carlo methodology and GB/SA theory was greatly appreciated throughout the research described in the first chapter of this dissertation. His eagerness to teach, troubleshoot, debug, and brainstorm

deserves the highest recognition, as does his uncanny ability to make even the most dry and technical aspects of chemical theory and software development seem exciting and provocative.

Other contributions from Drs. Sara Nichols and Laura DeFeo were appreciated as well, including early docking studies and insightful discussions about free-energy perturbation theory and sampling protocols. Some of the most fun and productive sessions I can recall were due in large part to their contribution of ideas, feedback, and undeniable wit. Indeed, my experience in lab was an enjoyable one thanks not only to their scholarly yet fun-loving attitudes but also the antics that made so many uneventful events so eventful.

My time at Yale has been memorable in large part due to the many friends that became part of my life. Drs. George Gardenier, Keith Whitener, and Rachel Dexter were perhaps most instrumental in keeping me sane outside of Sterling Chemistry Lab. Additional recognition goes to Mike Nordin for his always-cheerful greetings, as well as Kevin "Lazy sysadmin" Hart, Keith "We lost Keith again" Horne, Dan "Knew exactly what to do" Hendricks, Dustin "Don't mind me I'll just be in the corner" Barlow, Ken "is not impressed" Addison, and the rest of the crew for at the very least feigning interest in my research. All joking aside, I always knew I could count on my friends for whatever I needed. Thank you.

I would like to acknowledge my professors at the University of Redlands for their continued encouragement, support, and advice still many years after leaving Redlands. Their instruction and example during my undergraduate work has carried me through graduate school and has left an indelible impression on my academic career.

Most importantly, none of this would have been possible without the love and patience of my family. Their unrelenting belief in my ability to achieve an accomplishment like this was an inspiration, to say the least, and my immediate family, to whom this dissertation is dedicated, has been a constant source of strength during the final years of this Ph.D.

For Mom, Dad, and Erin.

Chapter 1

FEP/GBSA and the approximated generalized Born potential.

1.1 Introduction.

The first theoretical calculations in chemistry were performed in 1927 by Walter heitlon and Fritz London, two German physicists who had been studying new quantum mechanical treatments of exchange forces. Their landmark contributions to valence bond theory[1] brought chemistry for the first time under the umbrella of quantum mechanics and led to a new understanding of the nature of the chemical bond.[2] The potential role of computers in endeavors such as these did not become obvious until the 1940s and 1950s,[3] when emerging computer technology first allowed scientists to solve elaborate equations for complex atomic systems with an efficiency that had previously been unattainable by any other means. Since this advent, computers have come to play an essential role in nearly every area of chemistry, from the fundamental concepts of wave equations and configuration interaction to modern applications in reaction prediction, host–guest chemistry, and drug design.[4]

In a sense, it can be seen as fortuitous that the practical limitations of early computer systems forced investigators to study fundamental properties, interactions, and dynamics before being able to pursue more-complex questions. For instance, just as a correct prediction of the structure of the water dimer was necessitated before simulations of pure liquids could be realistically attained, so were the required computational resources available for one before the other. Indeed, owing to its ubiquity in nature, perhaps it is not surprising

that water has been one of the most exhaustively studied molecules[5–13] throughout the relatively short history of computational chemistry. Elegant in its structural simplicity while at the same time complex in its intermolecular dynamics, water remains a subject of intrigue for chemists interested in a myriad of topics ranging from simple interactions in isolated systems to complex dynamics in supramolecular or biological environments.

Despite the indispensable role of quantum mechanics in forging a sound theoretical basis for computational chemistry, human curiosity and scientific demand have quickly outgrown the limits of current computer technology. As such, it is unfeasible to tackle many of the most daunting challenges facing computational chemists today using solely methodologies as rigorous as quantum mechanics. This conclusion is easily illustrated: a molecule of medium size, containing 24 atoms, can be structure-optimized on a modern personal computer using state-of-the-art quantum mechanical theory in roughly 24 hours, but because the mathematical complexity and concomitant computational demand of such theory scales by N^6 (where N is the number of electrons in the system),[14] for a protein consisting of 10,000 atoms, a single optimization at the same level of theory would take 11.2 trillion years to complete.

Recognition of these scaling problems motivated chemists and physicists as early as the 1960s to pursue the development of other branches of chemical theory.[15–20] From these developments came molecular mechanics,[19,21] which uses classical Newtonian mechanics rather than quantum mechanics to describe molecular systems, modeling vibrating bonds as harmonic oscillators and defining dihedral torsions by a simple Fourier series. This simplified theory affords remarkable performance and the computational requirements scale much more linearly than quantum mechanical methods. Unfortunately, atoms and molecules in nature behave by the rules of quantum physics and not by those of Newtonian physics, and so it is not surprising that this increase in performance comes at the expense of theoretical accuracy. To curtail this loss of accuracy somewhat, most modern molecular mechanical parameters such as atomic charges, polarizabilities, van der Waals radii, and

force constants are derived from or fit to the results of experiment or sound quantum mechanical calculations.[21,22]

Owing to the speed at which these calculations can be performed for extremely large systems, simulations of pure liquids and of supramolecular or biological systems are particularly amenable to molecular mechanics. As such, the development of accurate, realistic water models based on molecular mechanical approaches has emerged as one of the most fervently pursued endeavors in computational chemistry.

1.1.1 Explicit and implicit water models.

Among the earliest water models were the SPC and SPC/E models of Berendsen[10] and the TIP models of Jorgensen.[8,9,11] The TIP models are perhaps the most widely adopted water models to date, and come with three-, four-, and five-point geometries (TIP3P, TIP4P, and TIP5P, respectively) to accurately reproduce the geometry of molecular water and the structure of liquid water, including properties like density and the temperature dependence thereof, heat of vaporization, and radial distribution functions.

These traditional models rely upon the explicit treatment of all atoms in the liquid; each atom is defined with its own coordinates, van der Waals radius, polarizability (if applicable), and net charge. All interactions between all atoms within their solvation shells or within a pre-defined cutoff are considered, and any change in the liquid structure as a result of a perturbation to the system necessitates that all interactions be relaxed, or equilibrated, to once again achieve the system's low-energy state. For a simulation of pure water consisting of 512 water molecules, this process is fairly trivial. However, for biomolecular systems wherein a protein is solvated by millions of water molecules and for which tens of millions of Monte Carlo moves or molecular dynamics steps are required to accurately simulate the desired process, sampling with explicit water requires significant computational time and resources. In a typical Monte Carlo simulation of a solute in water, the solute itself—ostensibly the component of primary interest—is sampled only

once every 60 moves; the other 59 moves are used to sample water configurations to minimize their contribution to the overall energy of the system. Using this sampling procedure, a prototypical molecular mechanical system of 3,000 solute atoms solvated by 12.5 million water molecules converges to a low-energy state after roughly 3 weeks of simulation time on a modern personal computer. When compared to the analogous gas-phase simulation for which only a few hours are required to achieve convergence, it becomes apparent that a more computationally feasible way of correctly capturing the effects of solvation is desirable. In addition to the slow convergence rates and long simulation times, often it is the case that explicitly treated water molecules become equilibrated in the binding site of a protein, complicating the system setup and affecting the outcome of the simulation.[23–25]

Despite the frustrations afforded by accounting for solvation, it remains an essential component of biomolecular simulations and its effects upon the system cannot be disregarded.

To circumvent several of these problems, "implicit" approaches to solvation, particularly solvation by water, have been developed. In an implicit solvation model, the effect of explicit solvent molecules is abstracted into a statistical dielectric continuum where the potential of mean force is applied to approximate the averaged behavior of the liquid;[26] implicit models are often referred to as "continuum" models for this most notable feature. This solution to the problem of solvation is particularly attractive because it simplifies the system being studied and can afford significant speed increases over explicit solvent models.

1.1.2 The GB/SA solvation model.

Of the many implicit solvent models available, those based upon the solvent-accessible surface area have arguably become the most popular because of their conceptual simplicity and ease of implementation in existing chemical modeling software; the free energy of

solvation is simply added to the potential energy U for the protein–ligand complex in the gas phase. Perhaps the most widely adopted flavor of the solvent-accessible surface area models is the generalized Born / surface area (GB/SA) model, formulated by Still and co-workers[27] in 1990 and revised[28] in 1997 as a fast, semi-analytical treatment of solvation for molecular mechanical simulations. In the GB/SA model, the free energy of solvation, G_{sol}, is given as the sum of three terms: G_{cav}, G_{vdW}, and G_{pol}, eq 1.1.

$$G_{\mathrm{sol}} = G_{\mathrm{cav}} + G_{\mathrm{vdW}} + G_{\mathrm{pol}} \tag{1.1}$$

In eq 1.1, G_{cav} is a solvent–solvent cavitation term describing the free energy required to form a cavity within the solvent to accommodate the shape and volume of the solute. G_{vdW} accounts for the solute–solvent van der Waals interactions, and G_{pol} describes solute–solvent electrostatic polarization. The G_{cav} and G_{vdW} terms are often grouped into a single nonpolar term, G_{NP}, which is linearly related[29] to the solvent-accessible surface area SA_i (Å^2) for each atom type i, eq 1.2. The σ_i (kcal/mol·Å^2) term in eq 1.2 is an empirically determined solvation parameter for cavity formation.

$$G_{\mathrm{NP}} = G_{\mathrm{cav}} + G_{\mathrm{vdW}} = \sum_i \sigma_i SA_i \tag{1.2}$$

As has been the case for many of the actively developed continuum solvent models, most of the effort that has gone into refining GB/SA theory has focused on improving the description of the electrostatic contribution.[26,30–32] From the original formulation of the GB/SA model,[27] G_{pol} is based on the Born equation for a spherical body with charge q and radius α in a solvent dielectric ε, eq 1.3. The electrostatic term is then expanded to allow for application to irregularly shaped solutes in the generalized Born equation, eq 1.4.

$$G_{\mathrm{pol}} = -166.0 \left(1 - \frac{1}{\varepsilon}\right) \frac{q^2}{\alpha} \tag{1.3}$$

$$G_{\text{pol}} = -166.0\left(1 - \frac{1}{\varepsilon}\right)\sum_i \sum_j \frac{q_i q_j}{r_{ij}^2 + \alpha_{ij}^2 \exp(-r_{ij}^2/2\alpha_{ij}^2)^{1/2}} \tag{1.4}$$

In eq 1.4, the electrostatic polarization term G_{pol} is determined by taking the sum of the electrostatic energies of all atom pairs. The electrostatic energy for a given atom pair is calculated as a function of the dielectric constant, ε, of the solvent being simulated, the charges associated with each atom in a pair, q_i and q_j, the distance between each atom in a pair, r_{ij}, and the Born radius associated with each atom in a pair, α_i and α_j. The Born radius α_i is taken as the spherically averaged distance from the center of atom i to its dielectric boundary, where $\alpha_{ij} = (\alpha_i \alpha_j)^{1/2}$, illustrated in Figure 1.1.

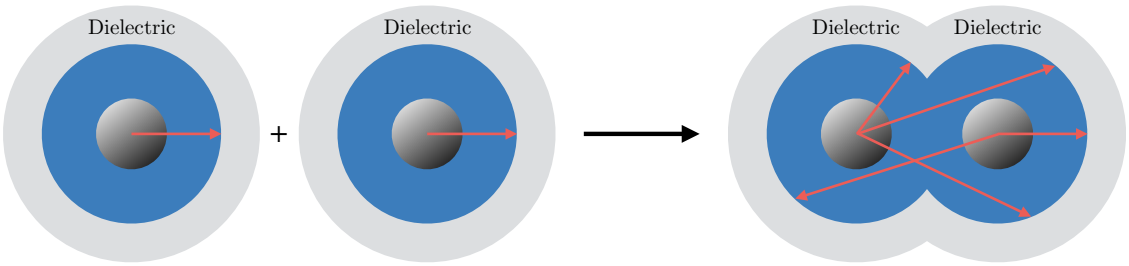


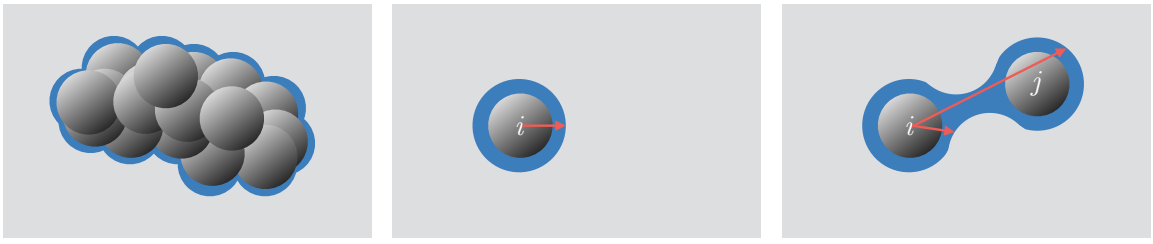
Figure 1.1: The pairwise nature of the Born radius.

The exponential function in eq 1.4 is exploited to force G_{pol} to approximate the dielectric part of Coulomb's law as atoms i and j move beyond the contact distance of their Born radii. It should be noted that solving eq 1.4 for G_{pol} is not trivial, owing to the pairwise nature of α_{ij}; computing α_{ij} itself requires a numerical finite-difference method, which, while yielding well-defined Born radii, becomes prohibitively time-consuming as the systems under investigation become increasingly complex.

In the 1997 revision to GB/SA,[28] the approach to calculating the Born radii was made fully analytical, circumventing the need to employ a numerical method to compute the electrostatic polarization term. This was accomplished by the definition of a new term, $G'_{\text{pol},i}$, eq 1.5.

$$G'_{\mathrm{pol},i} = \frac{-166.0}{R_{\mathrm{vdW},i} + \phi + P_1} + \sum^{1,2} \frac{P_2 V_j}{r_{ij}^4} + \sum^{1,3} \frac{P_3 V_j}{r_{ij}^4} + \sum^{1,4,\mathrm{NB}} \frac{P_4 V_j \mathrm{CCF}}{r_{ij}^4} \tag{1.5}$$

Thus, in eq 1.5, the partial electrostatic contribution of atom i is taken as the sum of self, 1,2, 1,3, 1,4, and 1,>4 terms relating to every other atom j. P_1–P_4 are optimized scaling factors for each respective interaction type, and CCF is a close-contact function for 1,4 and 1,>4 interactions. The dielectric offset, ϕ, defines a gap between the edge of the van der Waals radius of atom i and the beginning of the dielectric continuum, making the implicit solvent behave more like an explicit solvent. V_j is the volume occupied by atom j, and r_{ij} is the distance between atoms i and j. The concept is illustrated in Figure 1.2. Imagine a system of neutral atoms, as in (a), where $G_{\mathrm{pol}} = 0$. If one were to remove all the atoms except for one, and a charge were placed on that single atom, i, the system in (b) would then possess some nonzero G_{pol}, $G'_{\mathrm{pol},i}$. Using the Born equation (eq 1.3), one could then easily compute α_i. If one were to add a second atom j back into the system but keep j uncharged, then G_{pol} would change by V_j/r^4 due to displacement of the continuum by atom j. Finally, placing the charge back onto j, as in (c), would change G_{pol} by $G'_{\mathrm{pol},j}$. The procedure could be repeated for all other atom pairs until a final G_{pol} is achieved. This approach in principle makes the simulation of large biomolecular systems with GB/SA solvation much more feasible and has been widely accepted as the standard, modern form of GB/SA theory.[33,34]



(a) (b) (c)

Figure 1.2: Born radii are calculated using the analytical approximation to G_{pol}.

1.1.3 GB/SA solvation in protein simulations.

There is great appeal in employing implicit solvents in simulations of large biomolecular systems,[35] primarily due to the simplification they can afford over systems modeled in explicit solvent. In the last 10 years, GB/SA has been widely used in the context of molecular dynamics simulations, including studies of RNA hairpin unfolding,[36] nucleic acid conformational dynamics,[37] and determination of free-energy surfaces of β-hairpin and α-helical peptides by replica-exchange molecular dynamics.[38] There is also precedent for the use of GB/SA in Metropolis[39] Monte Carlo simulations of biomolecular systems. For instance, flexible docking[40] and concerted rotation with angles[41,42] algorithms take advantage of GB/SA solvation.

Unfortunately, the number of reported studies using GB/SA in Monte Carlo simulations of proteins pales in comparison to the number of reported studies using GB/SA in molecular dynamics simulations of proteins, and few have exploited GB/SA in rigorous calculations of binding affinities.[43–46] To understand this somewhat surprising finding, one must consider the impact of GB/SA theory on the methodological requirements of Monte Carlo vs. molecular dynamics simulations. Indeed, despite the simplification afforded by switching from an explicit to an implicit solvent model, simulations of large systems in implicit solvent can become similarly as computationally demanding as those in explicit solvent; whereas the computational demand and slow rate of convergence of modeling a system in explicit solvent arises from the large number of solvent molecules, high computational demand of modeling a system with GB/SA arises from the pairwise nature of the Born energy. For example, in a theoretical three-atom system consisting of atoms i, j, and k, the Born energy for atom pair ij will change if the position of atom k changes during a Monte Carlo move, because the Born radii of atoms i and j are both dependent on the position of atom k. Thus, in the context of a Monte Carlo simulation, one quickly finds that all atom-pair energies in a given system must be recalculated after every move, even if the position of most atoms in the system did not change. Accordingly, the sheer number of

energy calculations that must be performed in a large system is staggering. Furthermore, a Monte Carlo simulation requires significantly more moves than a molecular dynamics simulation requires time steps; it is therefore not surprising that molecular dynamics simulations in GB/SA exhibit only a 4–5-fold increase in computation time relative to the gas phase,[47] whereas we find that Monte Carlo simulations in GB/SA exhibit a 15–20-fold increase. In this light, the Monte Carlo method may quickly lose its appeal to those attempting to simulate large biomolecular systems with GB/SA. Nevertheless, the many powerful algorithms only available within a Monte Carlo manifold make it an indispensable tool for studying such systems.

Several modifications to the GB/SA model have been proposed in the literature.[48–51] Rather than focusing on the form of G_{pol}, these approaches have developed procedures to determine which energies need to be recalculated after every move and which ones do not. For instance, the "frozen atom" approximation of Still and co-workers[48] approximates the effects of atoms distant from the site of interest by freezing their coordinates throughout the duration of the simulation. Pairwise terms involving these frozen atoms therefore need to be calculated only once, and the derivatives thereof need not be calculated at all. In addition, a buffer region is defined, containing frozen atoms that are close enough to the site of interest to experience energetically significant changes to their Born radii as a result of the moving atoms. It quickly becomes apparent that given a setup such as this, different series of calculations can be performed for each atom pair based upon whether the pair is defined as frozen–frozen, moving–moving, buffer–frozen, or buffer–moving. This method was benchmarked using trial systems of camphor bound to cytochrome P450 and benzamidine bound to β-trypsin. Depending upon the system and the cutoff distance that was chosen, speed increases of approximately 1.5–10.6-fold were achieved.

A different approach was taken by Michel and co-workers[49] in 2006, in which they used the pairwise descreening approximation (PDA)[52] for their description of the Born radii and structured their GB/SA implementation in such a way that the energy of an

atom pair is only recalculated if the Born radius of either atom changes by more than a specified threshold after a Monte Carlo move. They determined the optimized threshold to be 0.005 Å, yielding results within less than 0.1% error of a fully rigorous calculation over the course of 5,000 moves. With this modification, a 3-fold speed increase was attained, and even with a threshold as low as 0.001 Å, a 2.5-fold speed increase was attained. In addition, when the simplified sampling potential proposed by Gelb[50] in 2003 was implemented alongside, the approximations afforded a 7–8-fold speed increase over a fully rigorous Monte Carlo GB/SA calculation.

1.1.4 Monte Carlo free-energy perturbation.

Free-energy perturbation (FEP) theory is a powerful method for calculating free-energy differences (ΔG) of two different chemical states, **A** and **B**. Based on statistical mechanics and the Zwanzig equation,[53] eq 1.6, it is used in both the molecular dynamics and Monte Carlo manifolds.

$$\Delta G(\mathbf{A} \longrightarrow \mathbf{B}) = G_{\mathbf{B}} - G_{\mathbf{A}} = -k_{\mathrm{B}}T \ln \left\langle \exp\left(-\frac{E_{\mathbf{B}} - E_{\mathbf{A}}}{k_{\mathrm{B}}T}\right)\right\rangle_{\mathbf{A}} \tag{1.6}$$

The difference in free energy between states **A** and **B** is calculated as an ensemble average of a simulation for state **A**. In practice, a Monte Carlo simulation is run for state **A** to determine $E_{\mathbf{A}}$, and each time a new configuration is accepted, the energy $E_{\mathbf{B}}$ is calculated for state **B** as well. For states that differ significantly, a scaling parameter λ is used to divide the simulation into a series of small windows, allowing the system to be perturbed smoothly over several simulations from $\lambda = 0$ at state **A** to $\lambda = 1$ at state **B**. Simulations can also be run in reverse, from $\lambda = 1$ at state **B** to $\lambda = 0$ at state **A**, and hysteresis between the forward and backward simulations can be minimized by employing various techniques; using the smallest possible increments of λ ensures smooth convergence and minimizes hysteresis. Typically, a doublewide sampling method is used, where an accepted configuration is perturbed to both $-\Delta\lambda$ and $+\Delta\lambda$ at once. Other sampling

methods have been proposed, including overlap and double-ended variations.[54] States **A** and **B** can differ in any number of ways, making FEP a very versatile computational tool. For instance, bond lengths or nonbonded interatomic distances can be perturbed from state **A** to **B**, yielding a potential energy-surface map along one or more sets of reaction coordinates. Alternately, atom types can be perturbed, simulating a theoretical mutation of molecule **A** into molecule **B**. The latter is particularly useful in determining relative free energies of hydration or relative free energies of binding. When doing so, the appropriate thermodynamic cycle is considered and two FEP calculations are performed independently. Examples are shown in Figure 1.3.

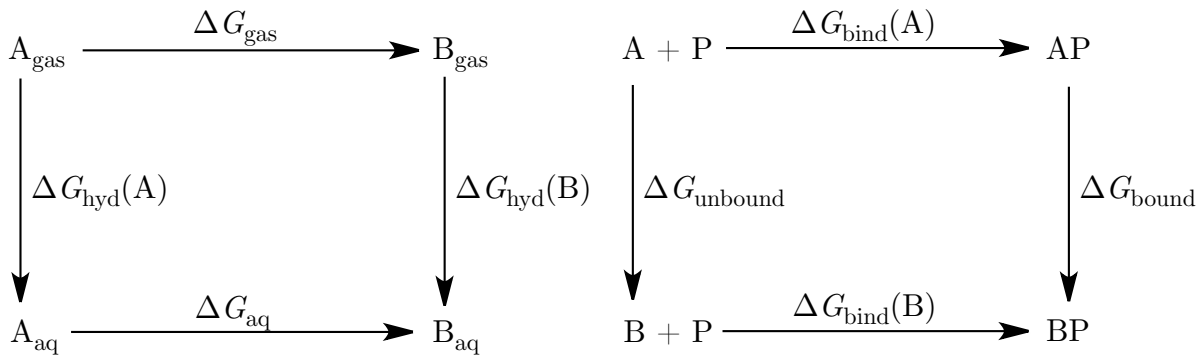


Figure 1.3: Thermodynamic cycles for FEP-based determination of free energies of hydration (left) and binding (right).

In Figure 1.3, the thermodynamic cycle on the left represents a FEP simulation for calculating the relative free energy of hydration, $\Delta\Delta G_{\mathrm{hyd}}$, between two molecules **A** and **B**. This quantity is determined in eq 1.7 by perturbing molecule **A** into molecule **B** in both the gas phase and in aqueous solution, affording values for $\Delta G_{\mathrm{gas}}(\mathbf{A} \longrightarrow \mathbf{B})$ and $\Delta G_{\mathrm{aq}}(\mathbf{A} \longrightarrow \mathbf{B})$, respectively.

$$\Delta G_{\mathrm{aq}}(\mathbf{A} \longrightarrow \mathbf{B}) - \Delta G_{\mathrm{gas}}(\mathbf{A} \longrightarrow \mathbf{B}) = \Delta G_{\mathrm{hyd}}(\mathbf{B}) - \Delta G_{\mathrm{hyd}}(\mathbf{A}) = \Delta\Delta G_{\mathrm{hyd}} \qquad (1.7)$$

Similarly in Figure 1.3, the thermodynamic cycle on the right depicts a FEP simulation for calculating relative free energies of binding, $\Delta\Delta G_{\mathrm{bind}}$, between two ligands **A** and **B**

to a protein **P**. This quantity is determined in eq 1.8 by perturbing ligand **A** into ligand **B** in both the unbound and bound states, affording values for $\Delta G_{\text{unbound}}(\mathbf{A} \longrightarrow \mathbf{B})$ and $\Delta G_{\text{bound}}(\mathbf{AP} \longrightarrow \mathbf{BP})$, respectively.

$$\Delta G_{\text{bound}}(\mathbf{AP} \rightarrow \mathbf{BP}) - \Delta G_{\text{unbound}}(\mathbf{A} \rightarrow \mathbf{B}) = \Delta G_{\text{bind}}(\mathbf{B}) - \Delta G_{\text{bind}}(\mathbf{A}) = \Delta\Delta G_{\text{bind}} \quad (1.8)$$

Of course, for realistic free energies of binding, solvation (typically by water) must be accounted for in both the unbound and bound states.

The implications of this theory to those the field of computer-aided drug design are exciting, to say the least; for any molecule **A** whose biological activity is known or well predicted, any number of modifications can be made and the relative biological activity for the derivative **B** can be easily predicted. Results can be corroborated by biological assay and further structural or chemical refinements can be made as necessary. Application of this theory can allow investigators to screen large sets of potential drug candidates with relative ease or to optimize an already-promising lead, making it more potent or perhaps altering its binding mode to target mutations in the host.[55]

1.1.5 Investigative focus.

The problem of correctly capturing the effects of solvation in applications of computer-aided drug design is not a trivial one, owing to large system size, structural complexity of solutes, mathematical obstacles, and stringent methodological requirements. Although continuum solvent models can aid in some of these aspects, they have historically been an incomplete solution to the problem; for example, the GB/SA solvation model is generally not amenable to FEP methods used in drug design, and specific details of implementation within a software package further complicate compatibility. Thus, we sought modification of the methodological approach to GB/SA calculations in the context of Monte Carlo free-energy perturbation. Our implementation is modeled after that of Michel, Taylor,

and Essex.[49] This approximation is based on the assumption that the impact of a moving atom on the Born radius of a distant atom is small, and therefore that a significant number of pairwise energy calculations can be omitted with little to no impact on the total energy change of the system after a Monte Carlo move. We sought to implement, test, and validate these modifications using real-world systems like those seen in our current drug-design projects, including benchmarking to document the performance of different subroutines and statistical analysis to fully understand the impact of our approximations.

1.2 Experimental design and results.

The scope of this work was three-fold. The first goal was to evaluate the existing implementation of GB/SA in terms of both performance and accuracy. The second goal was to modify the existing implementation of GB/SA to be compatible with existing FEP subroutines and to evaluate the performance and accuracy of both small and large FEP/GBSA simulations. Free energies of binding were to be compared to analogous ones in TIP4P explicit water. The third goal was to implement an approximated generalized Born potential, modeled after that of Michel, Taylor, and Essex.[49] The effects of the approximation were to be evaluated in terms of speed, accuracy, and ability to scale, again comparing free energies of binding to those of analogous TIP4P simulations. In all cases, care was taken to adhere to the appropriate standards of system setup, end-user involvement, and consideration of available computational resources.

In the search for non-nucleoside inhibitors of HIV-1 reverse transcriptase, lead optimization of structures that were generated from similarity searches have lead to active compounds consisting of a 1,3,4-oxadiazole core linking a phenyl and an anilinyl ring.[56,57] We chose as our test system a structure from an NNRTI series for HIV-RT, consisting of a parent ligand 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine and ten of its monochloro analogs bound to HIV-1 RT. The protein scoop contained 2,728 atoms in 178 residues, and the parent ligand contained 30 atoms. The parent ligand **1** is shown in Figure 1.4.

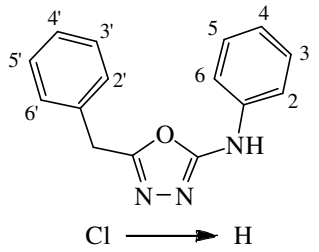


Figure 1.4: The parent ligand **1**, 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine, in our test system. Monochloro substitution has been investigated at the ten positions indicated.

Additional test cases used throughout this work consisted of two sets of small organic molecules involving perturbations commonly seen in FEP simulations of drug-like molecules, including a PhX $\longrightarrow$ PhY series, Tables 1.1 and 1.2. Unless otherwise noted, all calculations were performed using a modified version of *MCPRO*[58] version 2.05 on Linux.

1.2.1 Preliminary evaluation of GB/SA performance.

To evaluate the performance and accuracy of the unmodified version of GB/SA in *MCPRO*, standard Monte Carlo simulations were performed on our test system (**1**, Figure 1.4) bound and unbound, in the gas phase and with GB/SA, with 1.66×10^5 configurations of equilibration followed by 5.0×10^5 configurations of averaging for each of 14 windows of doublewide sampling. These values are scaled by 1/60 from those of a typical simulation in explicit water to account for the solute/solvent move ratio. Periodically, interactions of residues of similar charge were monitored; in our experience when using GB/SA in simulations of large proteins, the Coulombic forces between residues are often miscalculated if the GB/SA parameterization for the Born radii is not correct. For instance, two lysine residues might be brought together or even fused if the exaggeration is large enough. We envisioned that this could be a problem that would need to be addressed before proceeding with any further enhancements of the GB/SA model. Gratifyingly, our inspections revealed no such phenomena. Benchmark simulation times are given in Table 1.3.

Table 1.1: Test perturbations of small organic molecules.

$PhCH_3$	$\longrightarrow$	PhOH
$PhC(O)CH_3$	$\longrightarrow$	PhEt
Benzene	$\longrightarrow$	Pyridine
Cyclohexane	$\longrightarrow$	THP
THP	$\longrightarrow$	Dioxane
Piperidine	$\longrightarrow$	THP
$AcOCH_3$	$\longrightarrow$	AcOH
AcOH	$\longrightarrow$	Acetone
$AcNH_2$	$\longrightarrow$	AcOH
$N(CH_3)_3$	$\longrightarrow$	Acetone
$N(CH_3)_3$	$\longrightarrow$	$NH(CH_3)_2$
$NH(CH_3)_2$	$\longrightarrow$	CH_3NH_2
iPrCl	$\longrightarrow$	Propane
Propane	$\longrightarrow$	Ethane
Propane	$\longrightarrow$	*n*-PrBr

Table 1.2: Test perturbations featuring the PhX $\longrightarrow$ PhY motif.

PhBr	$\longrightarrow$	PhCl
$PhCH_3$	$\longrightarrow$	$PhCF_3$
PhCl	$\longrightarrow$	PhF
PhCl	$\longrightarrow$	Benzene
PhF	$\longrightarrow$	Benzene
PhCN	$\longrightarrow$	PhF
$PhC(O)CH_3$	$\longrightarrow$	$PhC(O)NH_2$
$PhC(O)CH_3$	$\longrightarrow$	$PhNO_2$
PhEt	$\longrightarrow$	$PhCH_3$
PhEt	$\longrightarrow$	$PhOCH_3$
PhPr	$\longrightarrow$	$BzOCH_3$
PhSH	$\longrightarrow$	PhOH
iPrPh	$\longrightarrow$	PhEt
$PhCH_3$	$\longrightarrow$	PhCl
$PhCH_3$	$\longrightarrow$	Benzene
$PhCH_3$	$\longrightarrow$	$PhNH_2$
$PhNO_2$	$\longrightarrow$	PhOH
$PhNHCH_3$	$\longrightarrow$	$PhNH_2$
PhOH	$\longrightarrow$	PhF
PhOH	$\longrightarrow$	$PhOCH_3$
$PhOCH_3$	$\longrightarrow$	PhOH
$PhSCH_3$	$\longrightarrow$	$PhOCH_3$

Table 1.3: Preliminary Monte Carlo benchmarking results for our test system.

Conditions	Simulation time (h)	$\Delta t_{\text{bound vs. unbound}}$	$\Delta t_{\text{GB/SA vs. gas}}$
Unbound, gas phase	0.43	–	–
Bound, gas phase	4.51	$10.5x$	–
Unbound, GB/SA	0.62	–	$1.44x$
Bound, GB/SA	76	$123x$	$16.8x$

Table 1.3 shows several pieces of data of which it was important to remain cognizant while proceeding. First, the gas-phase simulations took approximately 10.5 times longer to complete in the bound form than in the unbound form, giving a sense of the size of the bound system relative to the unbound. Not surprisingly, the efficiency of the GB/SA simulations deteriorated rapidly with increasing system size: the GB/SA calculations took 123 times longer to complete in the bound form in the unbound form. Perhaps more interesting is the comparison of gas-phase simulation times to GB/SA simulation times. These data indicate that for the unbound system (comprised of 30 atoms), GB/SA simulation times were roughly on par with those in the gas phase, costing a slowdown of only a few minutes. However, the simulation of the bound ligand **1** in GB/SA took nearly 17 times longer than it did in the gas phase, leaving ample room for improvement.

In the unmodified version of *MCPRO*, single-point energy calculations and standard Monte Carlo simulations (both of which are nonperturbing techniques) were the only tools available with GB/SA solvation. Nevertheless, energies could still be calculated that were in some ways analogous to those acquired via Monte Carlo FEP in TIP4P explicit water. For example, relative free energies of solvation could be determined by using single-point GB/SA energy calculations for the initial and final states (**A** and **B**, respectively), eq 1.9.

$$\Delta G_{\text{sol}}(\text{SP/GBSA},\mathbf{B}) - \Delta G_{\text{sol}}(\text{SP/GBSA},\mathbf{A}) = \Delta\Delta G_{\text{sol}}(\text{SP/GBSA}) \qquad (1.9)$$

The value of $\Delta\Delta G_{\text{sol}}$(SP/GBSA) is analogous to $\Delta\Delta G_{\text{sol}}$(TIP4P), which could be computed using our standard MC/FEP protocol with TIP4P explicit water, eq 1.10.

$$\Delta G(\text{FEP/TIP4P}) - \Delta G(\text{FEP/Vacuum}) = \Delta\Delta G_{\text{sol}}(\text{TIP4P}) \tag{1.10}$$

Although these two energies are arrived upon by very different algorithms (single-point energy calculations involve no sampling whatsoever), during the early stage of the investigation we were interested in whether or not the effect of the water sampling in MC/FEP could be captured in simple single-point energy calculations with GB/SA solvation. Thus, values of $\Delta\Delta G_{\text{sol}}$(SP/GBSA) and $\Delta\Delta G_{\text{sol}}$(TIP4P) were computed for the entire test set of 55 small- to mid-size organic molecules. For the MC/FEP calculations in TIP4P explicit water, 1.0×10^7 configurations of equilibration and 3.0×10^7 configurations of averaging were accumulated with solute moves attempted every 60 configurations. The doublewide sampling method was used over 14 windows. Calculations took between 7 and 55 times longer with TIP4P compared to those in the gas phase; CPU times for GB/SA single-point calculations were negligible. A correlation plot is shown in Figure 1.5.

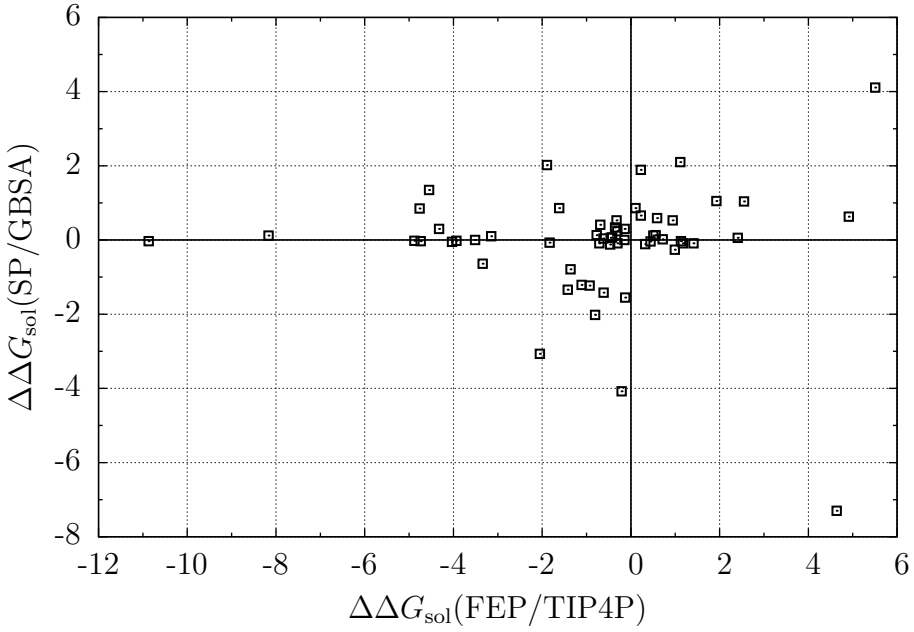


Figure 1.5: Relative free energies of solvation determined by SP with GB/SA and MC/FEP with TIP4P. Values are in kcal/mol. No correlation is observed ($r^2 = 0.0002$).

As is clear, there is essentially no correlation ($r^2 = 0.0002$) between the relative free energies of solvation computed via MC/FEP in TIP4P explicit water and those determined from single-point calculations with GB/SA, and similar results could be expected for relative free energies of binding. Therefore, as was anticipated, the tools currently available with GB/SA solvation were insufficient for applications in drug design and as such there was clear motivation to pursue modifications allowing the use of GB/SA in the context Monte Carlo free-energy perturbation.

1.2.2 Program structure and technical challenges.

The unmodified version of GB/SA as implemented in *MCPRO* was not compatible with MC/FEP simulations in part because it assumed a static approach to atomic properties, whereas in FEP, properties such as van der Waals radii, atomic volumes (minus any overlap with neighboring atoms), and Coulombic charges must be dynamically adjusted and scaled as perturbations are made. For example, in the unmodified GB/SA implementation, the charges for saturated alkanes were dealt with in a united-atom fashion, such that the charges of the alkane hydrogens were moved onto the attached alkane carbon. Thus, for methane, four formal charges of +1.09 corresponding to the four hydrogens were added onto the formal −4.36 charge on carbon, making all the atoms as a group Coulombically neutral. Although this type of simplification is often desirable, in a perturbation it can be logistically problematic. Take for instance the perturbation from methane to chloromethane, where a hydrogen atom is perturbed to a chlorine atom. In a united-atom approach with FEP, the charge on the perturbed hydrogen atom would need to be identified and smoothly moved off of the carbon, so that the charge would be fully condensed at the beginning of the perturbation but would be fully expanded, with the appropriate new charge and volume for chlorine, at the end of the perturbation. Switching from a united-atom approach to an all-atom approach would circumvent complications such at this, except that specific GB/SA parameters were originally developed with the

united-atom approach in mind. Furthermore, as originally implemented, contribution of 1,2 and 1,3 interactions to $G'_{\mathrm{pol},i}$ in eq 1.5 were counted for certain types of 1,2 or 1,3 neighbors and were ignored for others. In a FEP simulation, however, these contributions might need to be counted for a given atom pair at one end of the perturbation but not at the other. Again in the case of perturbing methane to chloromethane, 1,2 interactions would be ignored entirely for methane, but would need to be counted only for the C–Cl atom pair in chloromethane.

Yet another component of FEP calculations that complicates integration with GB/SA is management of dummy atoms, which are special atom types whose properties are not felt by any other part of the system. Since an atom might require dummy characteristics at one end of the perturbation but not at the other, the properties of such atoms would also need to be dynamically adjusted and scaled as perturbations are made. When accommodating special cases like these, care must be taken to ensure that the energy of the system is not only consistent at any given point during the perturbation, but also that the change in energy of the system is smooth throughout the perturbation. Thus, several novel modifications were made to the GB/SA module in *MCPRO* with the goal of making GB/SA calculations fully compatible with FEP simulations. Specific details follow.

A new subroutine, `GETSIG`, is called at the initialization of each window in order to identify atoms whose initial and/or final states require special treatment for the determination of their van der Waals radii when FEP is being performed. Prior GB/SA code had no knowledge of initial and final states, a concept necessitated by many aspects of FEP calculations. In the original implementation of GB/SA in *MCPRO*, atoms were assigned a van der Waals radius in one of two ways: either implicitly, based on their Lennard–Jones parameters ε and σ as given by eq 1.11, or explicitly, based on the identity of the atom. For instance, the van der Waals radius of a benzene carbon is implicitly calculated to be 1.775 Å based on its Lennard–Jones parameters, whereas the van der

Waals radius for a hydrogen on a heteroatom is explicitly assigned to 1.15 Å and the van der Waals radius for covalently bonded fluorine is explicitly assigned to 2.0 Å. The implicitly based assignments are fully compatible with FEP, since the Lennard–Jones parameters are received by the GB/SA module after having been scaled appropriately by the FEP code base in *MCPRO*. However, the explicit assignments are inherently incompatible with FEP since the identity of any given atom can change during an FEP run. Therefore, the `GETSIG` subroutine acts as a decision-making "switch" to decide where and how van der Waals radii are assigned. If an FEP simulation with GB/SA solvation is requested, the `GETSIG` subroutine checks initial and final atom types and explicitly assigns van der Waals radii to the appropriate atom types. Dummy atoms are treated explicitly at this point and are assigned a van der Waals radius of 1.15 Å. If an atom requires an explicitly assigned van der Waals radius at one end of the perturbation and an implicitly calculated van der Waals radius at the other end, then special action is required: the Lennard–Jones parameters are determined for the atom type whose van der Waals radius would be implicitly calculated, and the van der Waals radius is then calculated for the appropriate state and stored as an explicitly assigned value. At the end of the `GETSIG` subroutine, all initial and final van der Waals radii are passed to the `GBSASETUP` subroutine, where scaled radii are calculated for the reference, first and second perturbed states, given by eq 1.12.

$$R_{\mathrm{vdW}} = \frac{1}{2}\left(\frac{\sqrt{4\varepsilon\sigma^{12}}}{\sqrt{4\varepsilon\sigma^{6}}}\right)^{1/3} \tag{1.11}$$

$$R_{\mathrm{vdW}} = \lambda R_{\mathrm{vdW}}(\mathrm{Final}) + (1-\lambda) R_{\mathrm{vdW}}(\mathrm{Initial}) \tag{1.12}$$

Atoms whose van der Waals radii require no explicit assignment need no further treatment; the van der Waals radii for their reference, first and second perturbed states are calculated based solely on their pre-scaled Lennard–Jones parameters for the entirety of the simulation.

The `GBSASETUP` subroutine in the GB/SA module is called immediately after the `GETSIG` subroutine and is responsible for computing the first three terms of $G'_{\mathrm{pol},i}$ in eq 1.5. This series of calculations required several significant modifications, the first of which was to include in FEP simulations the proper treatment of the united-atom approach to Coulombic charge on saturated alkyl groups. To this end, there are two scenarios: the first, where charges must be condensed at the beginning of the simulation but not at the end, and the second, where charges must be condensed at the end of the simulation but not at the beginning. Accounting for this is trivially accomplished by using the same scaling factor λ used in the calculation of van der Waals radii.

Two examples can be used to illustrate the first scenario. First, in a perturbation from methane to chloromethane, the charge on the hydrogen that is perturbed to chlorine must be condensed in the case of methane but not in the case of chloromethane. Second, in a perturbation from methane to ammonia, the charge on all of the hydrogens must be condensed in the case of methane but not in the case of ammonia. Whereas the factors affecting charge condensation in the first example are determined by the identity of the hydrogen (atom i), in the second example they are determined by the identity of the carbon/nitrogen (atom j). In such events, scaling by λ is performed as given in eqs 1.13 and 1.14. Equations also apply for the first and second perturbed states.

$$q_{\mathrm{Born}}(i) = q_{\mathrm{Born}}(i) + (1-\lambda)q_{\mathrm{Born}}(j) \tag{1.13}$$

$$q_{\mathrm{Born}}(j) = \lambda q_{\mathrm{Born}}(j) \tag{1.14}$$

Thus, when the neighbor atom j exhibits 100% hydrogen-like characteristics (i.e., $\lambda = 0$), 100% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$ and 0% of $q_{\mathrm{Born}}(j)$ is retained. Similarly, when the neighbor exhibits 50% hydrogen-like characteristics (i.e., $\lambda = 0.5$), 50% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$, and 50% of $q_{\mathrm{Born}}(j)$ is retained, and when the neighbor exhibits 0% hydrogen-like characteristics (i.e., $\lambda = 1.0$), 0% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$, and 100% of $q_{\mathrm{Born}}(j)$ is retained.

For the second scenario, two possible examples also exist, which are essentially the reverse of the examples given in the first scenario. One could envision perturbation from chloromethane to methane, where charge condensation is not required for the chlorine atom at the beginning of the simulation but is required for the hydrogen atom to which the chlorine atom is perturbed at the end of the simulation. Additionally, one could envision perturbation from ammonia to methane, where charge condensation is not required for the hydrogen atoms at the beginning of the simulation but is required for the hydrogen atoms at the end of the simulation, since the identity of the nitrogen is perturbed from nitrogen to carbon. Scaling by λ is performed as given by eqs 1.15 and 1.16. Equations also apply for the first and second perturbed states.

$$q_{\mathrm{Born}}(i) = q_{\mathrm{Born}}(i) + \lambda q_{\mathrm{Born}}(j) \tag{1.15}$$

$$q_{\mathrm{Born}}(j) = (1 - \lambda) q_{\mathrm{Born}}(j) \tag{1.16}$$

Thus, when the neighbor exhibits 100% hydrogen-like characteristics (i.e., $\lambda = 0$), 0% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$ and 100% of $q_{\mathrm{Born}}(j)$ is retained. Similarly, when the neighbor exhibits 50% hydrogen-like characteristics (i.e., $\lambda = 0.5$), 50% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$, and 50% of $q_{\mathrm{Born}}(j)$ is retained, and when the neighbor exhibits 0% hydrogen-like characteristics (i.e., $\lambda = 1.0$), 100% of $q_{\mathrm{Born}}(j)$ is added onto $q_{\mathrm{Born}}(i)$, and 0% of $q_{\mathrm{Born}}(j)$ is retained.

Another modification requiring similar attention to detail was the Coulombic charge correction to N–O compounds. As standard procedure, in the event of an N–O group 24% of the oxygen's charge is added to the nitrogen and removed from the oxygen. Since the decision to make such a correction is dependent on nonstatic atom types, in the unmodified implementation of GB/SA in *MCPRO* this correction was not compatible with FEP. Simple decision-making statements were added to identify N–O pairs that are perturbed to non-N–O pairs, and in such events, the percent of the correction that is

made is scaled from 100% at the beginning of the simulation to 0% at the end of the simulation. Similarly, for non-N–O pairs that are perturbed to N–O pairs, the percent of the correction that is made is scaled from 0% at the beginning of the simulation to 100% at the end of the simulation. In modifying the GB/SA module to accommodate FEP in calculating atomic volumes, a significant challenge was encountered in correctly calculating volume overlap of adjacent atoms. Generally, the atomic volume for an atom i is calculated as the volume of a sphere with radius R_{vdW}, minus the sum of any volume overlap from 1,2 neighboring atoms j. However, for certain atom-type pairs (such as C–F in fluoromethane), this method of calculating overlap of two atoms i and j can yield negative overlap, eq 1.17.

$$\mathrm{overlap}(i) = (\pi/3)(R_{\mathrm{vdW}}(i)(1+\mathrm{ratio}))^2 \cdot (3R_{\mathrm{vdW}}(i) - R_{\mathrm{vdW}}(i)(1+\mathrm{ratio})) \qquad (1.17)$$

$$\text{where ratio} = \frac{R^2_{\mathrm{vdW}}(j) - R^2_{\mathrm{vdW}}(i) - r_0^2}{2R_{\mathrm{vdW}}(i) \cdot r_0}$$

For FEP calculations where both bond lengths and van der Waals radii change, the problem is exacerbated and volumes can become even more negative. Moreover, overlap of dummy atoms becomes problematic since they are given nonzero van der Waals radii. This was remedied through several adjustments: first, a maximum value of 1.000 was set for the ratio term, effectively restricting the perturbation so that an atom cannot be overlapped by a volume greater than its own. Second, the overlap for neighboring atoms that are perturbed to or from dummy atoms is scaled by λ such that 100% overlap is counted when the atom is not a dummy and 0% overlap is counted when the atom is a dummy. Lastly, a minimum volume is set to 1.000 $\mathrm{Å}^3$ after overlap for all atoms to prevent division by zero for atoms that are completely overlapped by neighbors.

Yet another aspect of the GB/SA implementation that becomes problematic in FEP calculations is the fact that 1,2 and 1,3 contributions to the electrostatic term are not counted for alkane C–H atom pairs; in the event of an FEP simulation where an alkane

C–H atom pair is perturbed in some way, the 1,2 and/or 1,3 contribution must be turned on or off. The initial appearance of the electrostatic term, $G'_{\mathrm{pol},i}$, is defined in eq 1.18 where $P_1 = 0.073$. This also applies for the first and second perturbed states. Note that dummy atoms, which were assigned a minimum R_{vdW} of 1.15 Å, acquire a nonzero $G'_{\mathrm{pol},i}$ at this stage, which is incorrect and will be addressed later in the calculation of Born energies.

$$G'_{\mathrm{pol},i} = \frac{-166.0}{0.09 \cdot R_{\mathrm{vdW}} \cdot P_1} \tag{1.18}$$

Several decision-making steps are then taken in order to correctly account for 1,2 and 1,3 interactions for certain atom-type pairs. Contribution of 1,2 interactions is ignored completely for hydrogen on CT/CY saturated carbons and for saturated carbons on hydrogen. Otherwise, full contribution of 1,2 interactions is given for the reference, first and second perturbed states by eq 1.19, where $P_2 = 0.921$.

$$G'_{\mathrm{pol},i} = G'_{\mathrm{pol},i} + (P_2 V_j)(1/r_0^4) \tag{1.19}$$

In the event of FEP, contribution of 1,2 interactions is ignored completely for the following pair transformations, where both the initial and final states include pairs for which 1,2 electrostatic contribution must be ignored.

C–H $\longrightarrow$ C–DUM
C–H $\longrightarrow$ DUM–C
H–C $\longrightarrow$ C–DUM
H–C $\longrightarrow$ DUM–C
C–DUM $\longrightarrow$ C–H
C–DUM $\longrightarrow$ H–C
DUM–C $\longrightarrow$ H–C
DUM–C $\longrightarrow$ C–H
N–DUM $\longrightarrow$ C–H/DUM
DUM–N $\longrightarrow$ H/DUM–C

Several cases exist where 1,2 electrostatic contribution must be scaled on throughout the course of an FEP simulation. Such pair transformations include, but are not limited to, the following, where the initial state involves non-1,2-contributing pairs and the final state includes 1,2-contributing pairs.

$$\begin{aligned} \text{C–H/DUM} &\longrightarrow \text{C–C} \\ \text{H/DUM–C} &\longrightarrow \text{C–C} \\ \text{N–DUM} &\longrightarrow \text{N–H} \\ \text{DUM–N} &\longrightarrow \text{H–N} \end{aligned}$$

Thus, 1,2 contribution for these pairs is scaled so that 0% is added at $\lambda = 0$ and 100% is added at $\lambda = 1$, for the reference, first and second perturbed states as in eq 1.20.

$$G'_{\text{pol},i} = G'_{\text{pol},i} + \lambda(P_2 V_j)(1/r_0^4) \tag{1.20}$$

Similarly, several cases exist where 1,2 electrostatic contribution must be scaled off throughout the course of an FEP simulation. Such pair transformations include, but are not limited to, the following, where the initial state involves 1,2-contributing pairs and the final state includes non-1,2-contributing pairs.

$$\begin{aligned} \text{C–C} &\longrightarrow \text{C–H/DUM} \\ \text{C–C} &\longrightarrow \text{H/DUM–C} \\ \text{N–H} &\longrightarrow \text{N–DUM} \\ \text{H–N} &\longrightarrow \text{DUM–N} \end{aligned}$$

Thus, 1,2 contribution by for these pairs is scaled so that 100% is added at $\lambda = 0$ and 0 is added at $\lambda = 1$, for the reference, first and second perturbed states as in eq 1.21.

$$G'_{\text{pol},i} = G'_{\text{pol},i} + (1 - \lambda)(P_2 V_j)(1/r_0^4) \tag{1.21}$$

Similar steps are taken for 1,3 electrostatic contribution. Again, 1,3 electrostatic

contribution is ignored for hydrogen on CT/CY saturated carbons and for saturated carbons on hydrogen; it is included fully for other cases where no FEP has been requested, and is scaled by λ for cases where 1,3 contribution must be turned on or off throughout the course of an FEP simulation. Such cases are identical to those for 1,2 contribution, and are given in eqs 1.22–1.24, where $P_3 = 6.211$. Equations apply for the first and second perturbed states as well. It should be noted that all `GETSIG` and `GBSASETUP` assignments and calculations up to this point are made once at the initialization of each window and are held constant throughout the remaining moves of that window; only the 1,4 and 1,>4 contributions to the energy are reevaluated after every move, since the 1,2 and 1,3 relationships (and thus the 1,2 and 1,3 electrostatic contributions) are unlikely to change within a given window.

$$G'_{\mathrm{pol},i} = G'_{\mathrm{pol},i} + (P_3 V_k)(1/r_0^4) \tag{1.22}$$

$$G'_{\mathrm{pol},i} = G'_{\mathrm{pol},i} + \lambda(P_3 V_k)(1/r_0^4) \tag{1.23}$$

$$G'_{\mathrm{pol},i} = G'_{\mathrm{pol},i} + (1-\lambda)(P_3 V_k)(1/r_0^4) \tag{1.24}$$

From this point, $G'_{\mathrm{pol},i}$ for the reference, first, and second perturbed states is passed to the `CALCEGB` subroutine, which is called after every move and evaluates the 1,4 and 1,>4 contributions to the energy. The close-contact function is evaluated for 1,>4 pairs and a Born factor term is established based on the close-contact function. The Born factor is then appropriately scaled by λ if either atom in the pair is a dummy atom at any point during an FEP simulation. Born radii are summed in a pairwise fashion over all atom pairs for the reference, first and second perturbed states as in eq 1.25.

$$\alpha_i = \alpha_i + \mathrm{bornfactor}(i,j)(V_j) \tag{1.25}$$

Thus, for pairs where a dummy atom is involved, the Born factor is scaled in such a way that its contribution to α_i is correct. The Born radii are subsequently updated as in

eq 1.26, and electrostatic pairs are defined as in eqs 1.27 and 1.28. Equations also apply for the first and second perturbed states.

$$\alpha_i = \frac{-166.0}{\alpha_i + G'_{\mathrm{pol},i}} \tag{1.26}$$

$$qq = q_{\mathrm{Born}}(i) \cdot q_{\mathrm{Born}}(j) \tag{1.27}$$

$$rr = \alpha_i \cdot \alpha_j \tag{1.28}$$

Note that in eq 1.27, qq is zero for pairs containing a dummy atom. Finally, several terms involving atomic coordinates are grouped into the energy components E_{pol} and F_{GB} as described by Still and co-workers,[28] and the energy E_{GB} of the whole system is calculated for the reference, first, and second perturbed states by eq 1.29.

$$E_{\mathrm{GB}} = E_{\mathrm{GB}} - E_{\mathrm{pol}} \left(\frac{\mathrm{fac}}{F_{\mathrm{GB}}} \right) \tag{1.29}$$

In eq 1.29, fac $= qq$ when $i \neq j$ or fac $= qq/2$ when $i = j$.

1.2.3 Test cases of typical perturbations.

1.2.3.1 GB/SA energy components.

In validating the new code implementing FEP with GB/SA, it was of utmost importance to verify that the total energy, energy components, and individual atomic properties were perturbed smoothly throughout a simulation; an irregular trajectory might indicate improper management of parameters or errors in the code. Accordingly, Monte Carlo free-energy perturbations were performed for all perturbations in the PhX $\longrightarrow$ PhY series (Table 1.2) using standard protocol in *MCPRO*, with GB/SA requested as a standard solvent. Atomic charges, volumes, and electrostatic contributions from key atoms were monitored, as was the total Born energy of the system, along the trajectory from $\lambda = 0 \longrightarrow \lambda = 1$. Simulations were run for 1.66×10^5 configurations of equilibration

followed by 5.0×10^5 configurations of averaging for each of 14 windows using doublewide sampling. Key atoms were chosen based on their role in the perturbation; in general, any atom that was perturbed or that was attached to a perturbed atom was considered. Examples below illustrate performance of the code in resolving the technical challenges highlighted in the previous section.

PhCl $\longrightarrow$ PhH. Results for the perturbation of chlorobenzene to benzene are given in Figure 1.6. Key atoms were determined to be the chlorine and the carbon to which the chlorine was attached. Given the aforementioned challenges with calculating van der

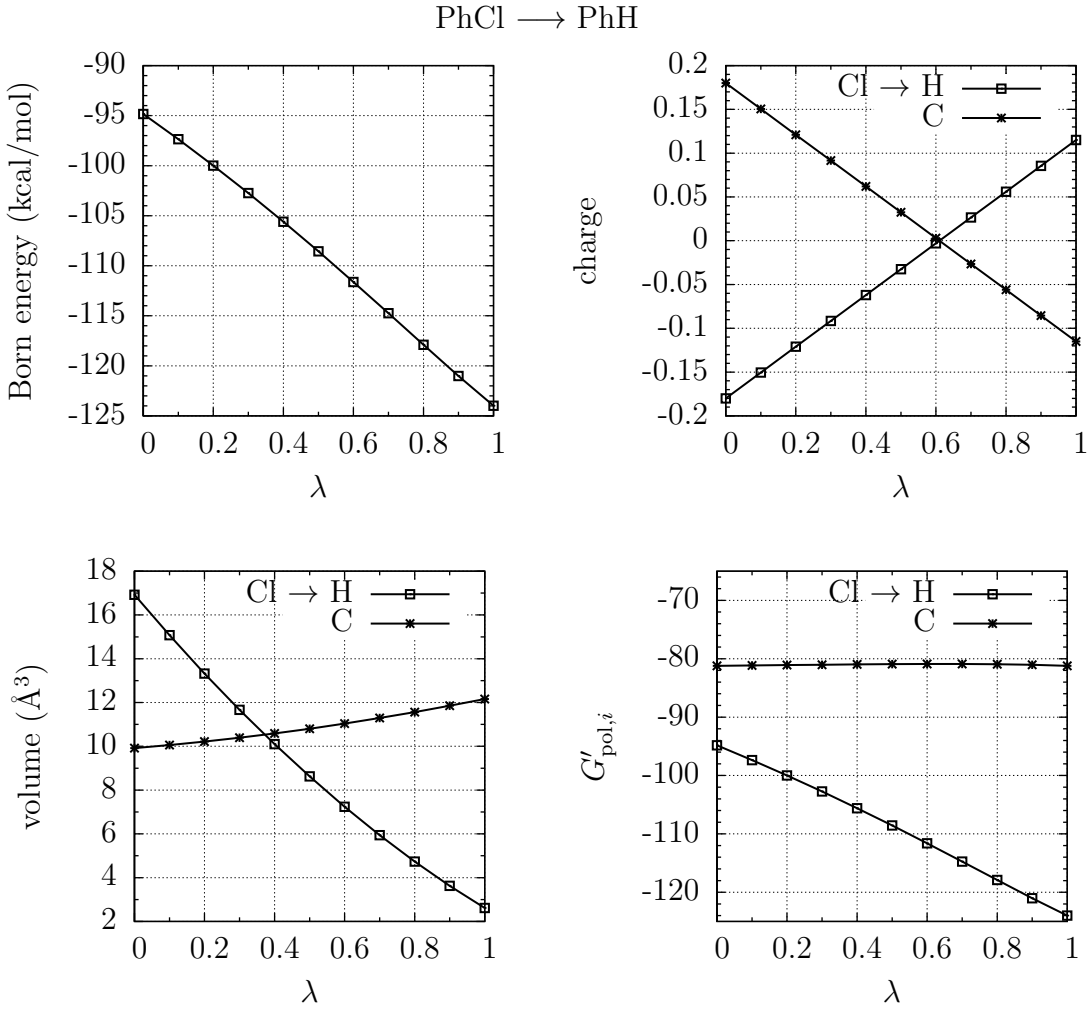


Figure 1.6: The Born energy and components thereof for the perturbation of chlorobenzene to benzene. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

Waals overlap in FEP, this example was particularly interesting as a test case because of the large differences in charge and volume between the initial and final state. All trajectories were smooth and consistent throughout the course of the perturbation. The charge trajectory (Figure 1.6, top right) reflected a reversal in charge in perturbing chlorine to hydrogen, consistent with the charge distribution observed for benzene. A similar reversal was seen in the case of atomic volumes, Figure 1.6, bottom left. A dramatic decrease in atomic volume was observed for the perturbation of chlorine to hydrogen, with concomitant increase in effective atomic volume for the attached carbon, owing to a decrease in van der Waals overlap. Electrostatic contribution (Figure 1.6, bottom right) also behaved as expected, and the resulting total Born energy followed a smooth arch with a maximum change in energy at $\lambda = 0.5$ and an overall ΔG_{Born} of +0.124 kcal/mol for the perturbation.

PhCN $\longrightarrow$ PhF. Results for the perturbation of cyanobenzene to fluorobenzene are given in Figure 1.7. Key atoms were determined to be the nitrogen being perturbed to a dummy atom, the carbon being perturbed to fluorine, and the carbon to which the carbon/fluorine was attached. Again, for all energies considered and components thereof, trajectories were smooth and consistent throughout the perturbation. This example was challenging as a test case due to the inclusion of a dummy atom, whose properties (or lack thereof) required special handling. Overall, the system was well behaved despite the challenges present. Of note were the trajectories of charge, volume, and electrostatic contribution for the nitrogen as it was perturbed to the dummy atom. Charge (Figure 1.7, top right) was scaled linearly from -0.43 for nitrogen to zero for the dummy; atomic volume (Figure 1.7, bottom left) was scaled linearly as well until a minimum of 1.0 Å^3 was achieved at $\lambda = 0.6$. In this case, the minimum volume was reached before $\lambda = 1.0$ because of additional volume loss from overlap of the adjacent perturbation of carbon to fluorine, also reflected in the figure. Lastly, effective electrostatic contribution of the nitrogen was scaled smoothly, albeit not entirely linearly, to zero for the dummy. Other

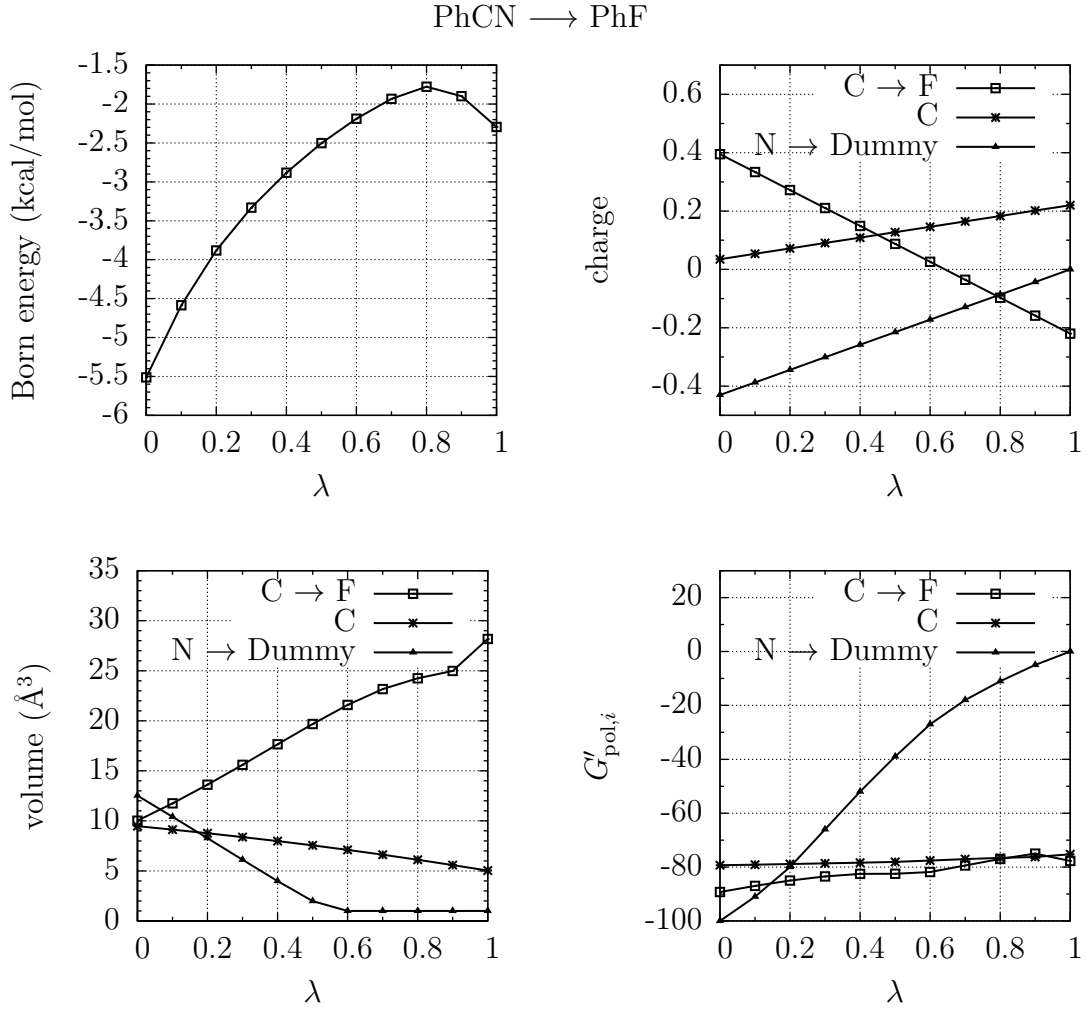


Figure 1.7: The Born energy and components thereof for the perturbation of cyanobenzene to fluorobenzene. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

transformations were executed as anticipated, and the overall Born energy followed a smooth arch with a maximum change in energy at $\lambda = 0.8$ and an overall ΔG_{Born} of +3.2 kcal/mol for the perturbation.

$\mathrm{PhC(O)CH_3} \longrightarrow \mathrm{PhC(O)NH_2}$. Results for the perturbation of acetophenone to benzamide are given in Figure 1.8. This test case was particularly complicated, featuring three perturbations (hydrogen on carbon to hydrogen on nitrogen, carbon to nitrogen, and hydrogen to dummy), which included a dummy atom as well as a geometry change (methyl to amino). Further complicating the energetics was the methyl contribution to

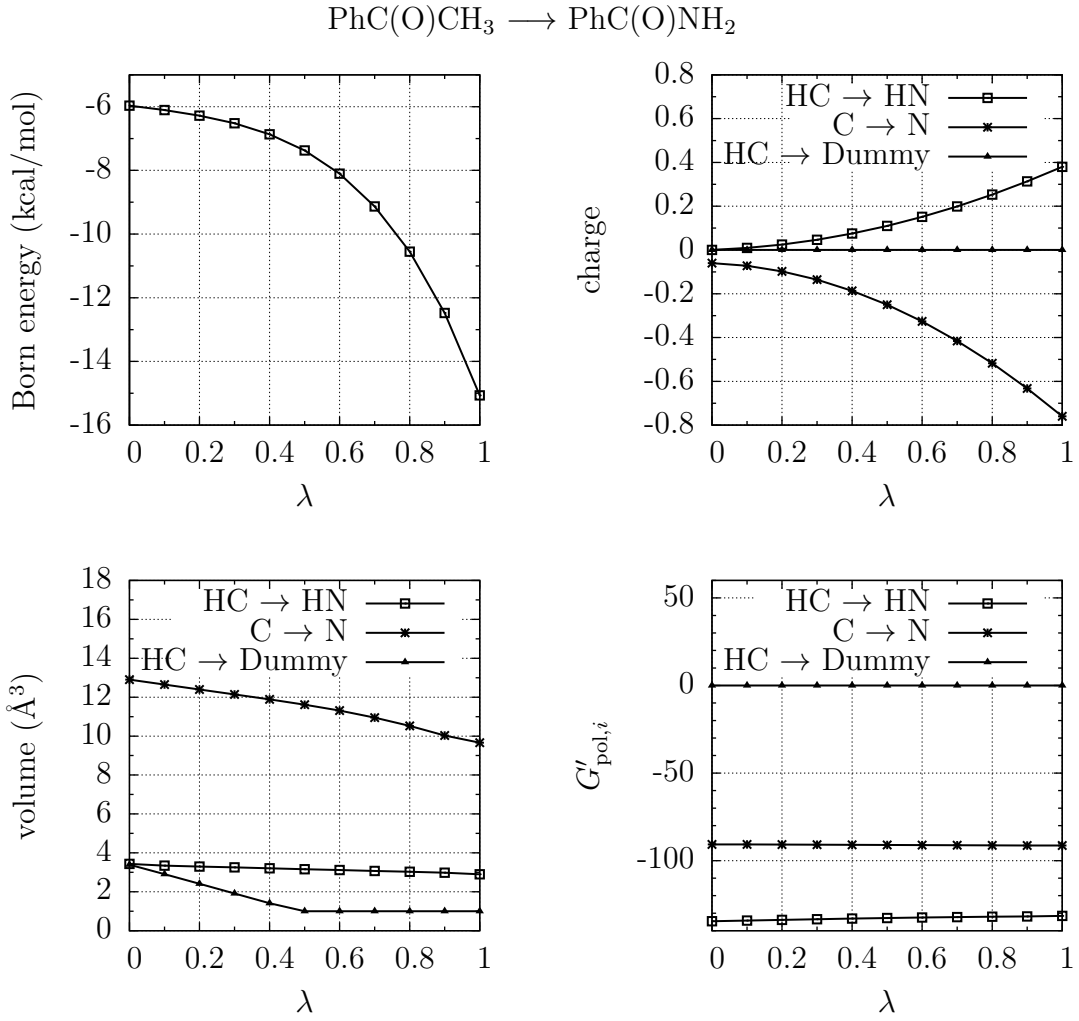


Figure 1.8: The Born energy and components thereof for the perturbation of acetophenone to benzamide. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

the electrostatic energy; recall that for H–C neighboring pairs where C is fully saturated, 1,2 and 1,3 contribution to the electrostatic energy is ignored. Thus, in this case, such contribution to the electrostatic energy is ignored for the methyl group in acetophenone but not for the amido group in benzamide. Gratifyingly, for all energies considered and components thereof, trajectories were smooth and consistent throughout the perturbation and trends played out as was anticipated. For atomic charges (Figure 1.8, top right), a significant change in charge was observed for methyl carbon to amido nitrogen. Of note was the observed change in charge for hydrogen on carbon to hydrogen on nitrogen. Recall

that hydrogens on saturated carbons are treated in a united-atom fashion such that their charges are condensed onto the neighboring carbon. Contrarily, hydrogens on heteroatoms are treated in an all-atom fashion. Thus, charges on the methyl hydrogens were expanded throughout the simulation, from zero at $\lambda = 0$ to +7.5 at $\lambda = 1.0$. However, for one methyl hydrogen this was not the case; the condensed-charge methyl hydrogen that was perturbed to a dummy atom maintained its zero charge throughout the perturbation. Schematically, the charge trajectories give a visual depiction to the overall charge dispersion for the perturbation of acetophenone to benzamide. Like in the perturbation of cyanobenzene to fluorobenzene, the hydrogen to dummy perturbation reached a minimum volume of 1.0 Å^3 near the midpoint of λ (Figure 1.8, bottom left) and the effective contribution to the electrostatic polarization of the methyl hydrogen to dummy was held at zero. Other transformations were unremarkable, yielding a smooth overall Born energy trajectory and ΔG_{Born} of −9.2 kcal/mol for the perturbation.

PhPr $\longrightarrow$ BzOCH_3 Results for the perturbation of propylbenzene to benzyl methyl ether are given in Figure 1.9. Key atoms were determined for this test case to be the carbon perturbed to oxygen, the methyl carbon, and the hydrogens perturbed to dummy atoms. This test case was notable in that it involved oxygen and simultaneous perturbation to two vicinal dummy atoms. Nevertheless, trajectories were smooth and consistent throughout the course of the perturbation for the total Born energy of the system, atomic charges, volumes, and electrostatic contributions, and trends therein were unremarkable. Volumes (Figure 1.9, bottom left) were stable despite vicinal hydrogen to dummy perturbations, with dummy atoms reaching the minimum volume of 1.0 Å^3 near $\lambda = 0.7$ and electrostatic contribution for the methylene hydrogens to dummy remaining constant at zero. The perturbation yielded an overall Born energy change ΔG_{Born} of −2.1 kcal/mol.

PhCH_3 $\longrightarrow$ PhCF_3. Results for the perturbation of toluene to (trifluoromethyl)benzene are given in Figure 1.10. Key atoms were determined to be the methyl hydrogens perturbed to fluorines and the carbon to which they were attached, as well as the substituted aromatic

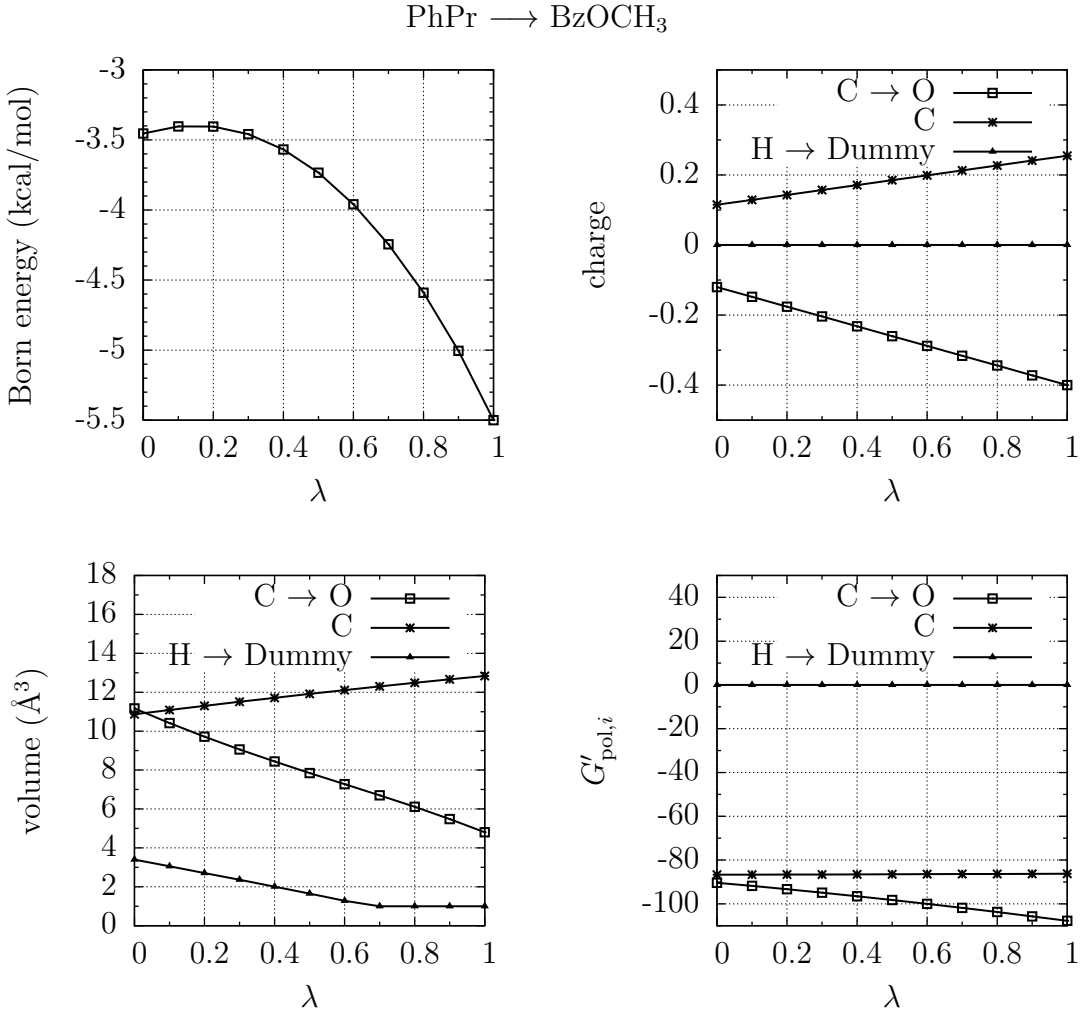


Figure 1.9: The Born energy and components thereof for the perturbation of propylbenzene to benzyl methyl ether. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

carbon. Once more, this example was particularly interesting as a test case because of the large differences in charge and volume between the initial and final state. Illustrating this is the volume trajectories graph (Figure 1.10, bottom left), which shows how the volume of the methyl carbon quickly vanished by $\lambda = 0.7$ as it became eclipsed by the three growing fluorines. Volume of the substituted aromatic carbon, the atom of the interest farthest from the fluorines, remained unchanged. Charge perturbations (Figure 1.10, top right) were large, most dramatically for the methyl carbon, which developed significant positive charge throughout the course of the simulation, owing to the large change in

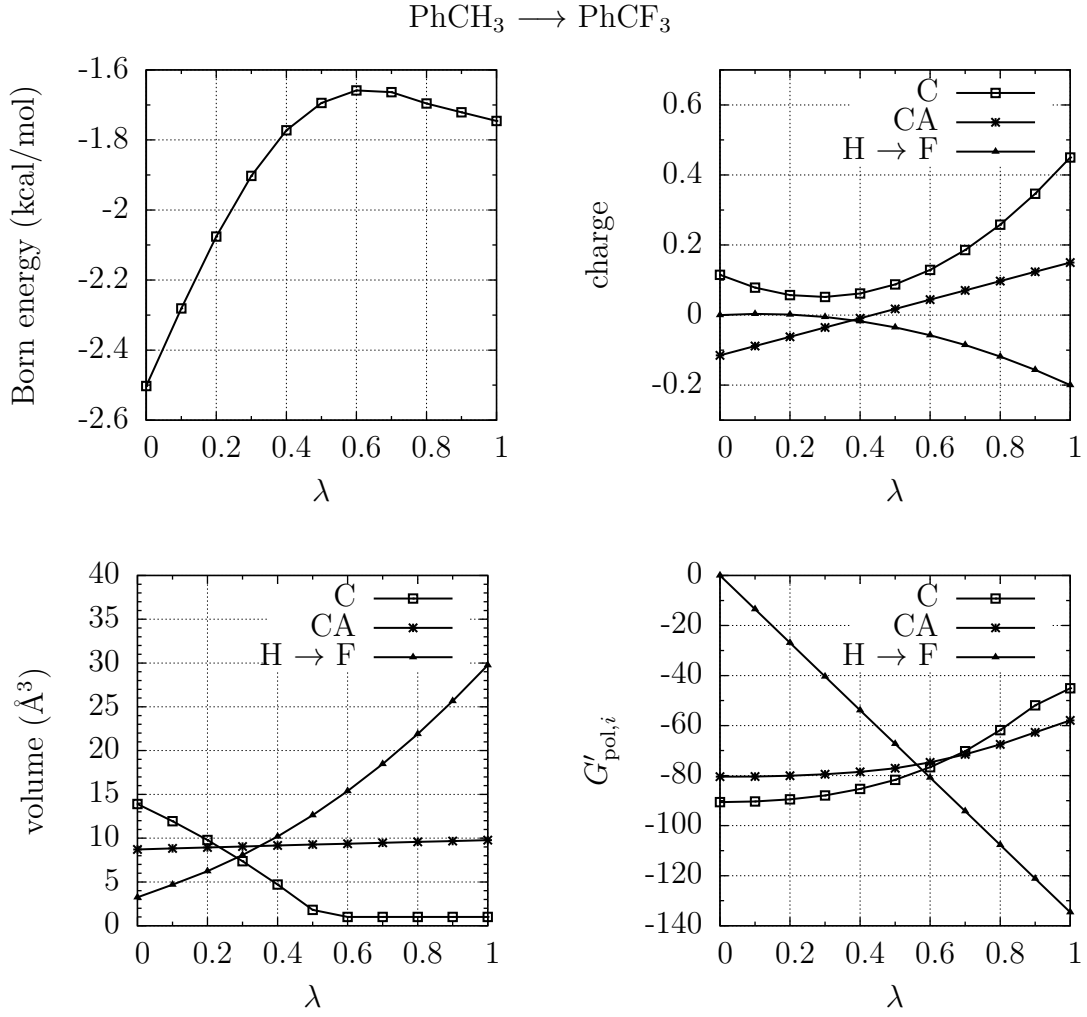


Figure 1.10: The Born energy and components thereof for the perturbation of toluene to (trifluoromethyl)benzene. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

electronegativity between methyl and trifluoromethyl groups. Electrostatic contribution for methyl hydrogens was scaled linearly from zero, yielding a smooth overall Born energy trajectory and ΔG_{Born} of +0.75 kcal/mol for the perturbation.

$\text{PhC(O)CH}_3 \longrightarrow \text{PhNO}_2$. Results for the perturbation of acetophenone to nitrobenzene are given in Figure 1.11. This test case was particularly complicated, featuring three perturbations (methyl carbon to oxygen, sp^2 carbon to nitrogen, and hydrogen to dummy) and evolution of a net charge. Further complicating management of charge was perturbation into two N–O atom pairs, a condition that necessitated special accommodation in

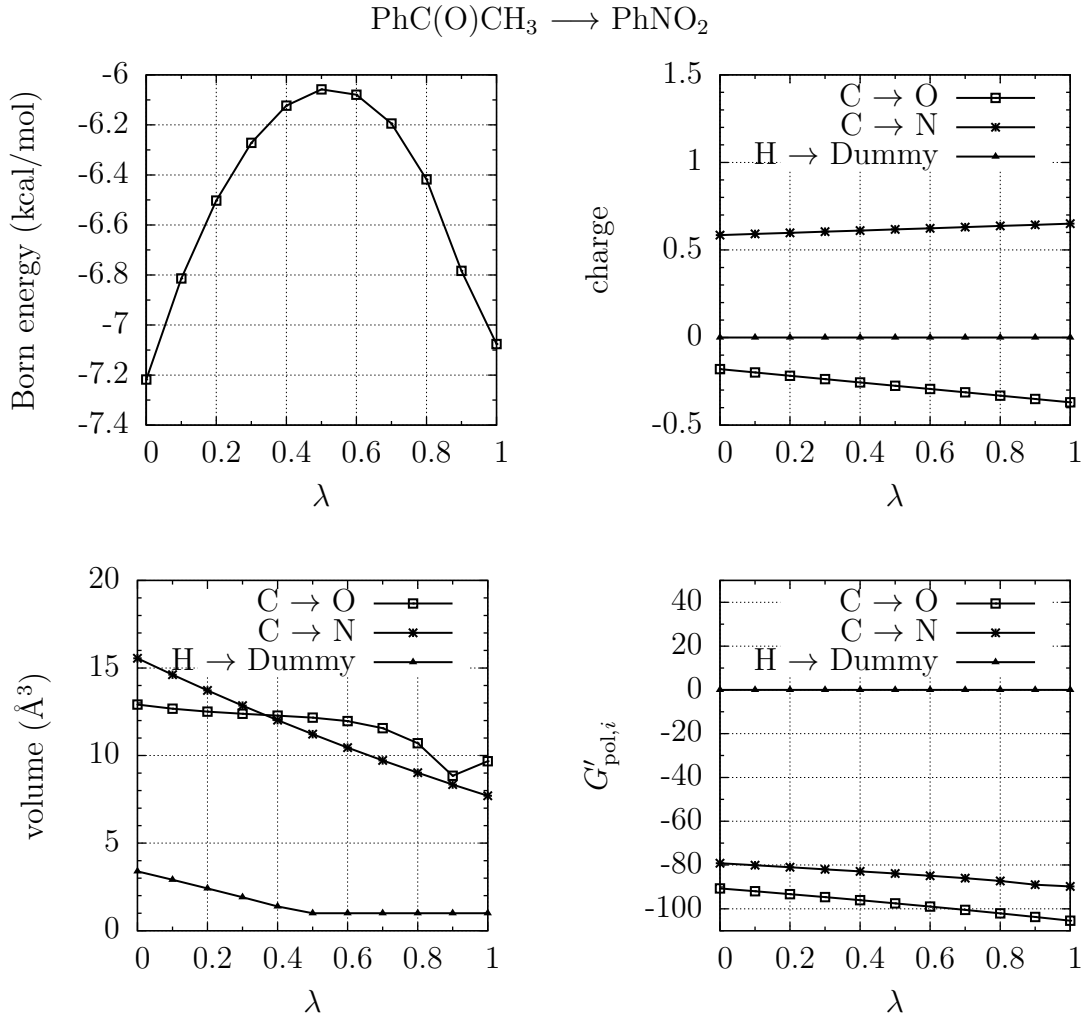


Figure 1.11: The Born energy and components thereof for the perturbation of acetophenone to nitrobenzene. Top left: energy trajectory; top right: charge trajectories; bottom left: volume trajectories; bottom right: electrostatic trajectories.

the FEP GB/SA code. In the event of an N–O group 24% of the oxygen's charge is added to the nitrogen and removed from the oxygen; in this case, the percent charge added to the nitrogen to each oxygen was scaled on throughout the perturbation, from 0% at $\lambda = 0$ to 24% at $\lambda = 1$. Gratifyingly, atomic charges (Figure 1.11, top right) were perturbed linearly throughout the simulation, with the charge on the methyl carbon becoming more negative as it was perturbed to oxygen, the charge on the sp^2 carbon becoming more positive as it was perturbed to nitrogen, and the charge on the methyl hydrogens and dummies remaining zero throughout. Volumes (Figure 1.11, bottom left) were predictable

with the exception of a small dip in volume for the methyl carbon to oxygen perturbation at $\lambda = 0.9$, although the trajectory was resolved at $\lambda = 1$. Dummy atoms reached the minimum volume of 1.0 Å^3 at $\lambda = 0.5$. Perturbation of electrostatic contribution (Figure 1.11, bottom right) was unremarkable, with contribution from methylene hydrogens to dummy remaining constant at zero. A smooth overall Born energy trajectory arch yielded a modest ΔG_{Born} of +0.15 kcal/mol for the perturbation.

These examples highlight the ability of our FEP implementation to handle special cases of charge united-atom to all-atom conversion including expansion/condensation and adjustment, as well as volume overlap, dummy atom and C–H contribution to $G'_{\text{pol},i}$, and geometric perturbations for a wide range of commonly encountered functional groups and functional group transformations.

1.2.3.2 Free energies of solvation.

Having confirmed that the energies and components thereof were well behaved for a wide range of functional group transformations, we were interested in investigating the performance of our implementation if GB/SA with FEP in computing free energies of solvation. Monte Carlo free-energy perturbations were performed for all transformations in Tables 1.1 and 1.2 in the gas phase, with GB/SA solvation, and with TIP4P explicit solvation. Standard setup was used in *MCPRO* with 14 windows of doublewide sampling. For the GB/SA and gas-phase calculations, simulations were run for 1.66×10^5 configurations of equilibration followed by 5.0×10^5 configurations of averaging per window, and for the TIP4P calculations, simulations were run for 1.0×10^7 configurations of equilibration followed by 3.0×10^7 configurations of averaging with solute moves attempted every 60 configurations. Using the concept illustrated in the thermodynamic cycle in Figure 1.3, free energies of solvation were then determined. Trajectories of select examples are presented. For each set of trajectories (one set containing GB/SA, TIP4P, and gas phase trajectories), free energies of solvation were determined by evaluating the energy

difference between each aqueous trajectory and the gas phase trajectory at their respective endpoints. Whereas only this endpoint was necessary in computing the free energies of solvation, examining the trajectories as a whole gave a sense of the evolution of the energies throughout the perturbation. Remarkably it was demonstrated that the GB/SA simulations yielded not only endpoints similar to those with TIP4P, but in most cases, the energies followed similar trajectories as well, Figures 1.12 and 1.13.

For the perturbation of chlorobenzene to benzene, $\Delta G_{\mathrm{GB/SA}}(\mathbf{A} \longrightarrow \mathbf{B}) = 1.104$ kcal/mol, $\Delta G_{\mathrm{TIP4P}}(\mathbf{A} \longrightarrow \mathbf{B}) = 0.904$ kcal/mol, and $\Delta G_{\mathrm{gas}}(\mathbf{A} \longrightarrow \mathbf{B}) = 1.261$ kcal/mol. Thus, relative free energies of solvation for the perturbation of chlorobenzene to benzene were -0.157 kcal/mol for GB/SA and -0.357 kcal/mol for TIP4P. For the PhX $\longrightarrow$ PhY set in general, GB/SA trajectories behaved more like gas-phase trajectories than TIP4P trajectories; GB/SA and gas-phase trajectories typically followed a more predictable path, whereas TIP4P trajectories often fluctuated, owing to variations in water molecule configurations. Such examples are given by chlorobenzene to benzene, cyanobenzene to fluorobenzene ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 3.325 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 2.408 kcal/mol), ethylbenzene to anisole ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -0.501 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 0.717 kcal/mol), as well as anisole to phenol ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -3.335 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -4.549 kcal/mol). Fortunately, most deviations were small and eventually corrected themselves. Additionally, there were many cases where the GB/SA and TIP4P trajectories mimicked each other's, illustrated by perturbations of acetophenone to benzamide ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -8.480 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -7.865 kcal/mol), propylbenzene to benzyl methyl ether ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -1.765 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -1.833 kcal/mol), thiophenol to phenol ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -4.758 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -4.764 kcal/mol), thioanisole to anisole ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 0.182 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 0.102 kcal/mol), as well as phenol to fluorobenzene ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 5.261 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 5.506 kcal/mol), and ethylbenzene to anisole ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -0.477 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 0.710 kcal/mol)

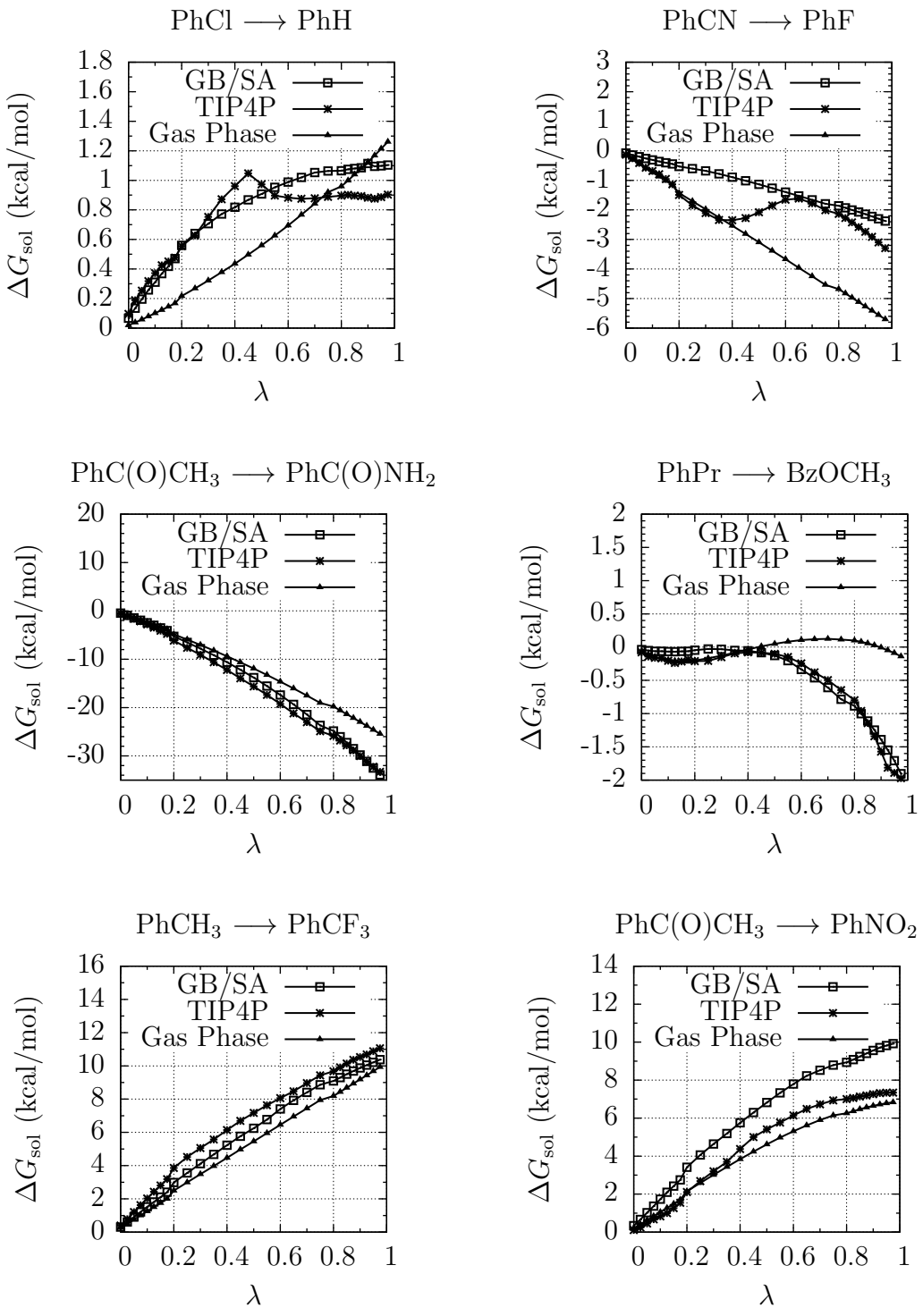


Figure 1.12: Selected trajectories for ΔG_{sol} from the PhX $\longrightarrow$ PhY (Table 1.2) series.

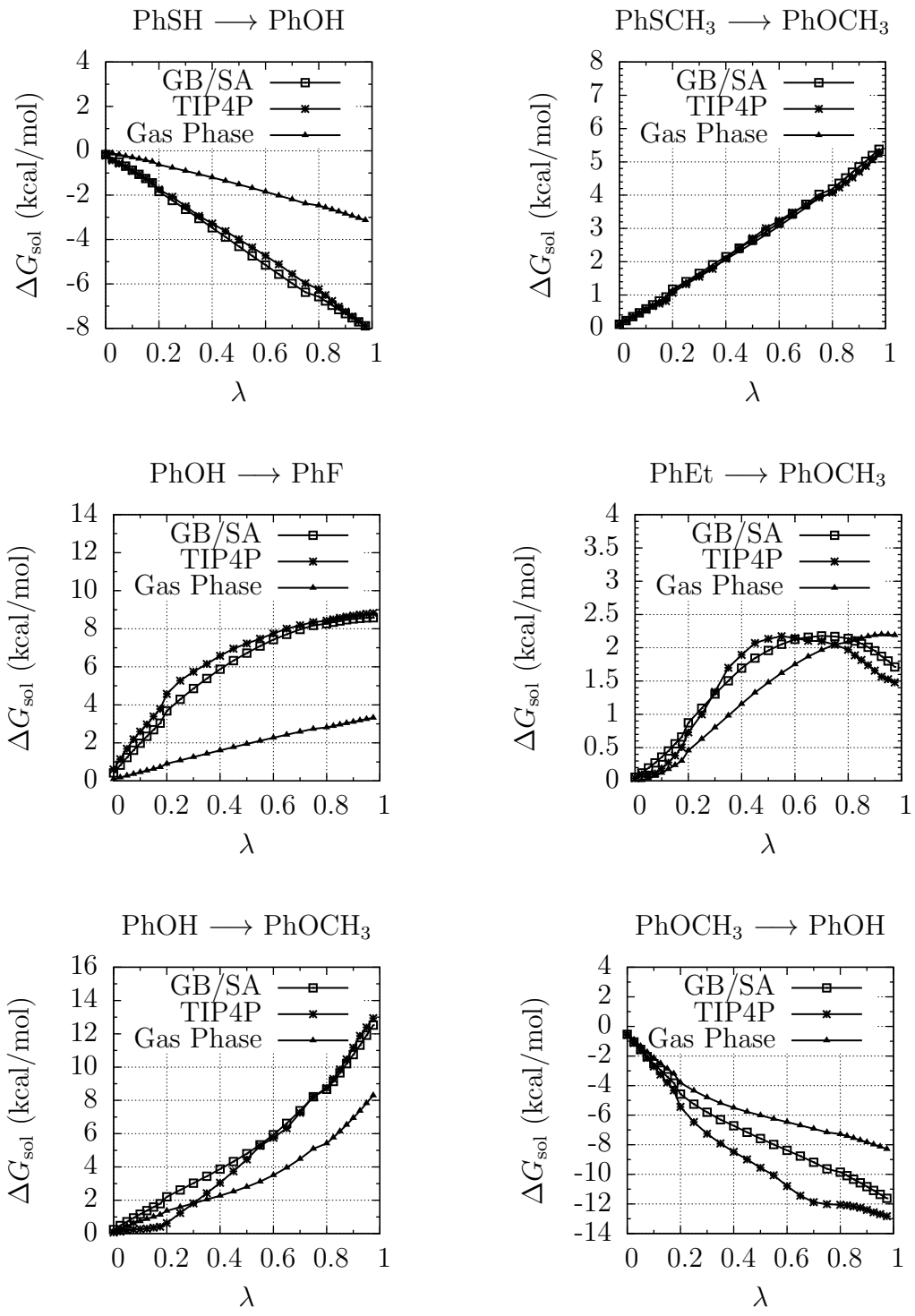


Figure 1.13: More selected trajectories for ΔG_{sol} from the PhX $\longrightarrow$ PhY (Table 1.2) series.

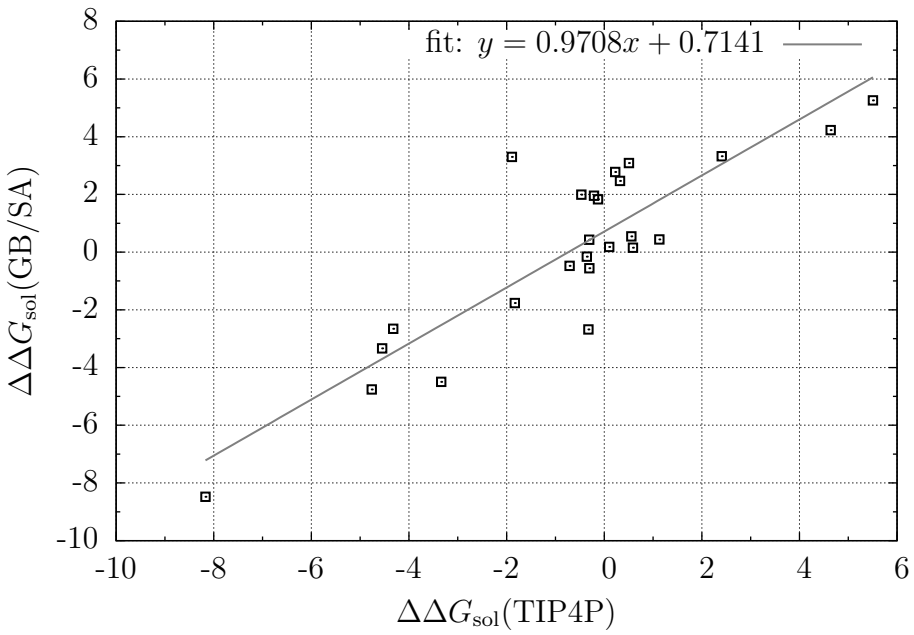


Figure 1.14: Correlation plot between GB/SA and TIP4P for the entire PhX $\longrightarrow$ PhY test set, $r^2 = 0.7588$, $y = 0.9708x + 0.7141$.

and phenol to anisole ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 4.227 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 4.640 kcal/mol). One example of a perturbation run forwards and backwards is given for phenol to anisole and anisole to phenol. Hysteresis was larger for GB/SA (forwards $\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 4.227 kcal/mol, backwards $\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -3.335 kcal/mol, hysteresis = 0.892 kcal/mol) than for TIP4P (forwards $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = 4.640 kcal/mol, backwards $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -4.549 kcal/mol, hysteresis = 0.091 kcal/mol). Overall, GB/SA performed reasonably well in reproducing relative free energies of solvation in TIP4P explicit water (Figure 1.14) with a correlation value $r^2 = 0.7588$.

Select results from the small perturbations test set illustrate similar trends, Figures 1.15 and 1.16. Again, GB/SA trajectories behaved more like gas-phase trajectories than TIP4P trajectories; with TIP4P trajectories often fluctuating and GB/SA and gas-phase trajectories following a more predictable path. Examples are given by ethane to methanol ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = 1.956 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) = -0.208 kcal/mol), cyclohexane to piperidine ($\Delta\Delta G_{\mathrm{sol}}$(GB/SA) = -4.499 kcal/mol, $\Delta\Delta G_{\mathrm{sol}}$(TIP4P) =

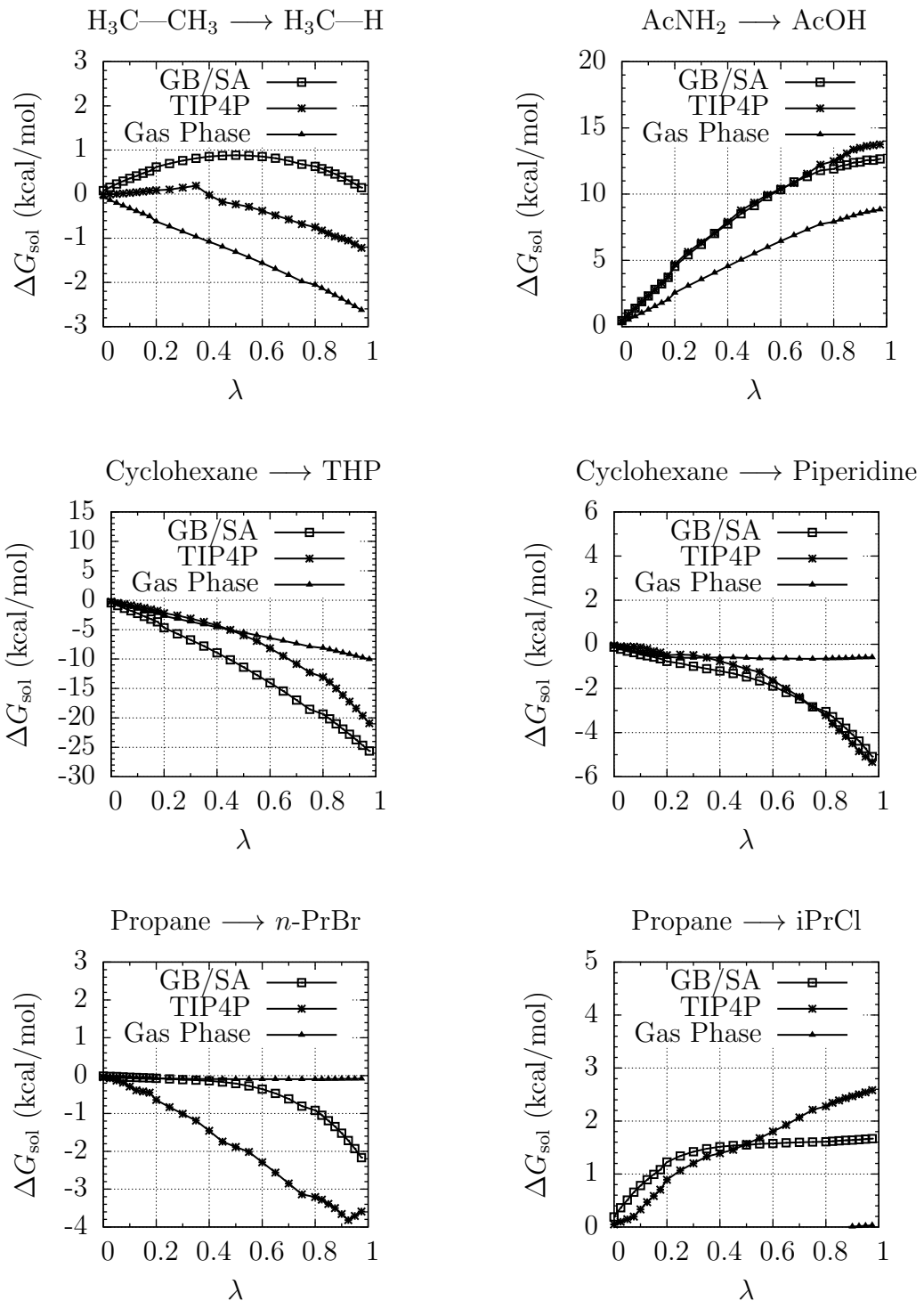


Figure 1.15: Selected trajectories for ΔG_{sol} from the small perturbations (Table 1.1) series.

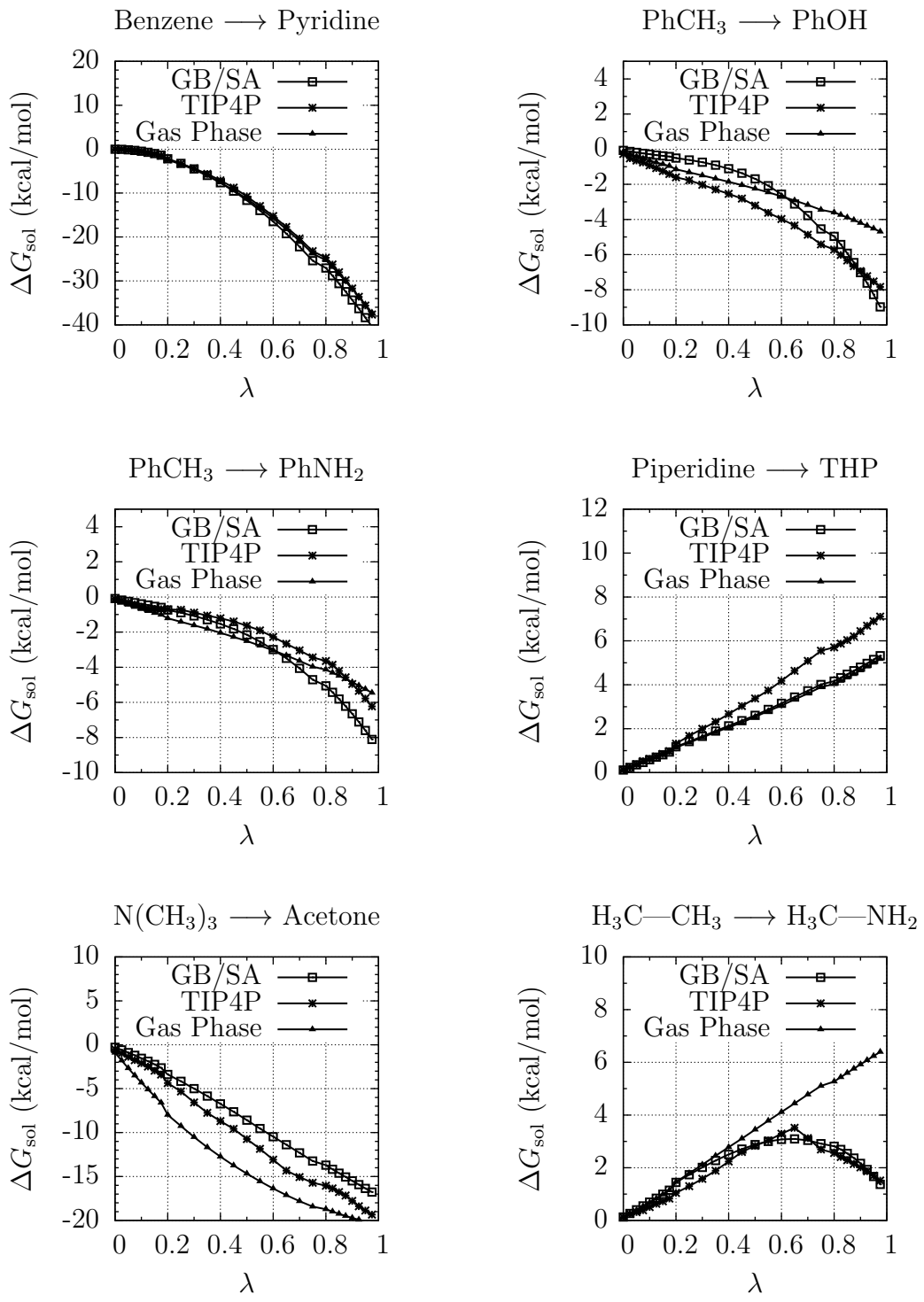


Figure 1.16: More selected trajectories for ΔG_{sol} from the small perturbations (Table 1.1) series.

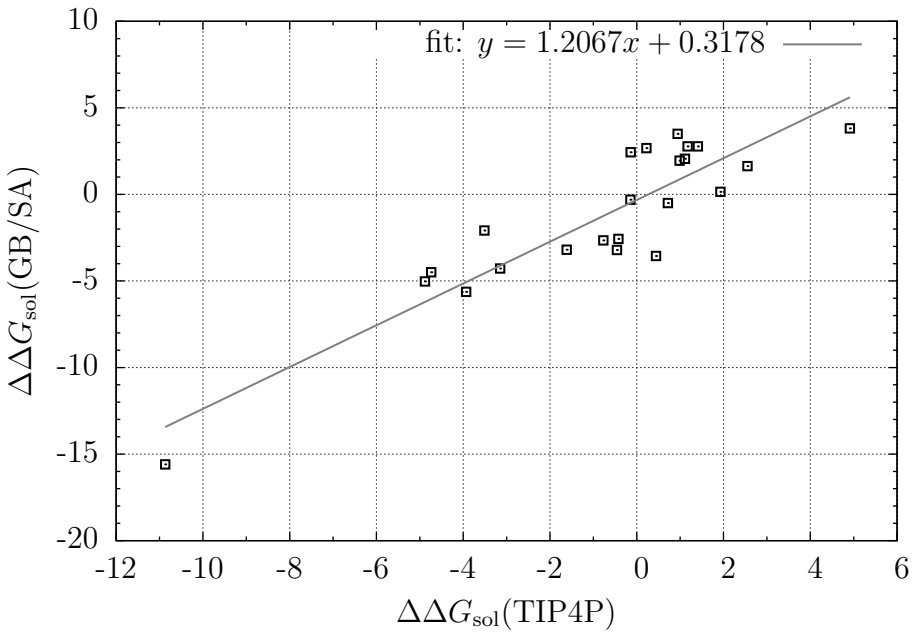


Figure 1.17: Correlation plot between GB/SA and TIP4P for the entire small perturbations test set, $r^2 = 0.8098$, $y = 1.2067x - 0.3178$.

-4.739 kcal/mol), propane to 1-bromopropane ($\Delta\Delta G_{\rm sol}$(GB/SA) = -2.086 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = -3.511 kcal/mol), 2-chloropropane to propane ($\Delta\Delta G_{\rm sol}$(GB/SA) = 1.637 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = 2.551 kcal/mol), and trimethylamine to acetone ($\Delta\Delta G_{\rm sol}$(GB/SA) = 3.50954 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = 0.941 kcal/mol). Cases where the GB/SA and TIP4P trajectories mimicked each other's were perturbations of acetamide to acetic acid ($\Delta\Delta G_{\rm sol}$(GB/SA) = 3.815 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = 4.911 kcal/mol), cyclohexane to piperidine ($\Delta\Delta G_{\rm sol}$(GB/SA) = -4.499 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = -4.739 kcal/mol), and ethane to methylamine ($\Delta\Delta G_{\rm sol}$(GB/SA) = -5.029 kcal/mol, $\Delta\Delta G_{\rm sol}$(TIP4P) = -4.881 kcal/mol). Again, GB/SA performed reasonably well in reproducing relative free energies of solvation in TIP4P explicit water (Figure 1.17) with a correlation value of $r^2 = 0.8098$. A correlation plot for all perturbations in our test set (combined PhX $\longrightarrow$ PhY and small perturbations) is given in Figure 1.18, showing a combined correlation value of $r^2 = 0.7642$.

Results in Figures 1.12–1.18 confirm that relative free energies of solvation can be well

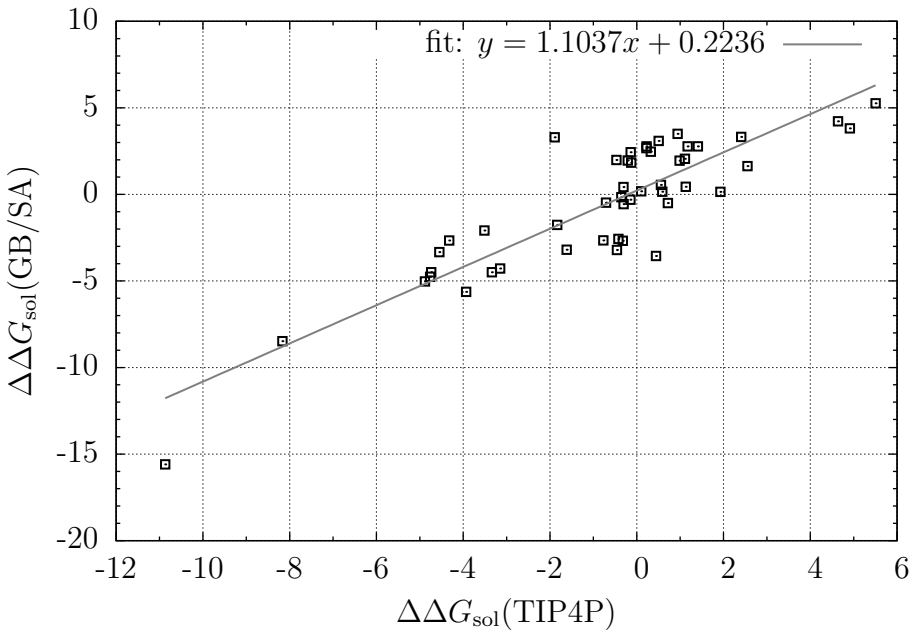


Figure 1.18: Correlation plot between GB/SA and TIP4P for all test perturbations, $r^2 = 0.7642$, $y = 1.1037x + 0.2236$.

estimated using FEP with GB/SA, and comparing correlation plots in Figures 1.14, 1.17, and 1.18 to Figure 1.5 confirms that FEP is desirable over single-point calculations as a method for accurately reproducing such energies. For these calculations, GB/SA afforded reasonable simulation times, given that the molecules were of average size. In our test set, GB/SA calculations took only 1.5 times longer than the same calculations in the gas phase, whereas calculations with TIP4P took approximately 53 times longer. In this respect, GB/SA is in fact very useable and may be considered an attractive alternative to explicit water models. However, the primary motivation for the implementation of GB/SA with FEP was for use in calculation of free energies of binding, and such calculations require FEP in the binding site of a protein. This becomes problematic because GB/SA does not scale well in large, biologically relevant systems without even considering the additional computational demand of FEP, as illustrated in Table 1.3. Therefore, we sought to implement an approximation to the generalized Born potential to improve performance of GB/SA in MC/FEP of large systems.

1.2.4 The approximated generalized Born potential.

Modeled after that of Michel, Taylor, and Essex,[49] our approximation to the generalized Born potential is based on early observations that in large systems, after any given GB/SA Monte Carlo move a substantial percentage of atoms experience a change in their Born radii on the order of only 10^{-10} Å, often much less. It is therefore conceivable that the impact of these pairs on the change in total energy of the system from one move to the next is negligible, and consequently that these pair energies in fact need not be updated after every move. Thus, we structured our implementation of GB/SA in *MCPRO* in such a way that the Born energy between an unmoving pair of atoms is only recalculated after a move if the Born radius of either atom has changed by more than a specified threshold, τ, since the last accepted move. In the event that this condition is not met, the Born energy for that pair is simply restored from the last accepted ensemble.

This approach is illustrated schematically in Figure 1.19, where shaded components comprise the unapproximated algorithm and unshaded components represent new components in the approximation. A simulation begins in the top left. After the first Monte Carlo move is made, Born radii are calculated analytically for all atoms and the electrostatic contribution for all atom pairs is determined. When Born radii are not computed numerically, this step is historically the most time consuming for large systems and without any approximations it is repeated for every move. After the Born energy of the entire system is determined, the energy is submitted to a Metropolis Monte Carlo acceptance test. If the move is accepted, Born radii for all atoms and the electrostatic contributions for all atom pairs are stored in memory to be used later; if the move is rejected, no information is retained. Another Monte Carlo move is then made and new Born radii are calculated. If no previous moves have been accepted, the process is repeated as if it were the first move. If, however, a previous move has been accepted, then two conditions are evaluated for each atom pair. The first evaluation is whether either atom in given pair has moved since the last accepted move. The second evaluation is whether the

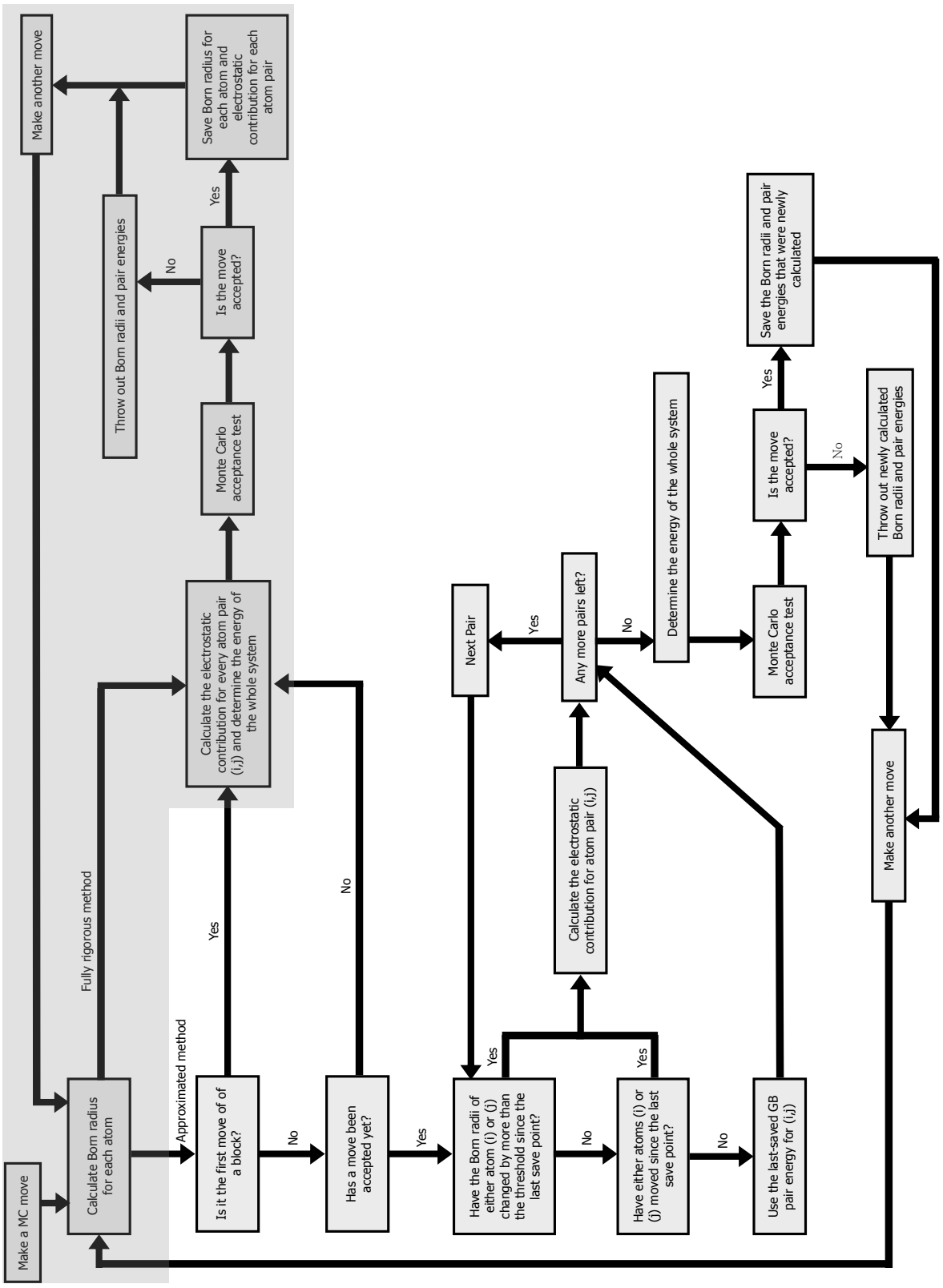


Figure 1.19: Schematic illustration of the approximated generalized Born algorithm. Shaded components comprise the unapproximated algorithm.

Born radius of each atom in a given pair has changed by more than the threshold τ since the last accepted move. If either of these conditions is met for either atom in the pair, then the Born energy for that pair must be recalculated because it is likely to have been significant to the energy of the system. However, if neither of the conditions are met for both atoms in the pair, then the energy calculation is skipped for that pair and instead is assigned from the corresponding pair energy from the last accepted move. In doing so, a large number of calculations can be skipped, affording an approximation to the total Born energy per move and throughout the ensemble. Once evaluations are made for every atom and atom pair, the Born energy of the entire system is determined and submitted to the Metropolis Monte Carlo acceptance test. If the move is accepted, updated Born radii and pair energies are stored in memory, whereas no information is retained if the move is rejected, and another Monte Carlo move is made. It is important to note at this point that the approximation is not (and cannot be) applied until a move is first accepted, and as such, a fully rigorous calculation set is performed on newly initialized runs for several moves, even when a threshold has been specified. This caveat is more significant in short simulations than long ones, where a larger percent of total moves (and thus a larger percent of unapproximated energy calculations) are made before the approximation can be applied.

1.2.5 Statistical significance of the approximation.

To illustrate the impact of the threshold on the number of atom-pair energy calculations skipped, FEP simulations were run for the C4 monochloro test perturbation of 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine (**1**, Figure 1.4) bound to HIV-RT at different threshold values. Values of τ ranging from 10^{-20} Å to 0.1 Å were selected and simulations were aborted once a move was accepted. Results are given in Figure 1.20. Our test system contained 2,758 atoms, comprised of 2,728 atoms in 178 residues in the protein scoop 30 atoms in the ligand, for a total of 3,803,282 unique atom pairs. Figure 1.20 illustrates

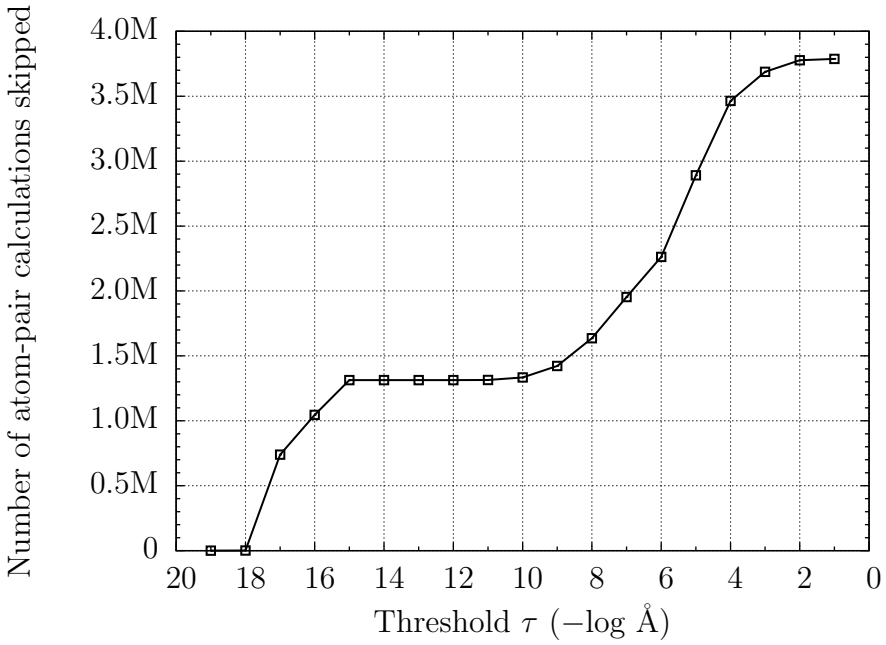


Figure 1.20: The number of atom-pair calculations skipped using our approximated potential, showing the impact of the threshold τ on the number of atom-pair energy calculations skipped for the C4 monochloro perturbation of **1** bound to HIV-RT.

that indeed a very large number calculations can be skipped for any given move, even at very small values of τ; at $\tau = 10^{-15}$ Å, 34% of atom-pair energy calculations (1,313,261) were skipped during a move and at $\tau = 10^{-4}$ Å, 91% of atom-pair energy calculations (3,464,227) were skipped during a move. As τ approached 1 Å, nearly all atom-pair energy calculations (99.5%, 3,786,992) were skipped once a full energy calculation had been accepted. Conversely, at values of τ near 10^{-18} Å, as few as 0.03% (1,195) atom-pair energy calculations were skipped.

Thus, the utility of τ serves not only to apply an approximation to the method by which total energies are evaluated, but also to allow the investigator to "dial in" as much or as little approximation as is desired. Although the exact behavior of τ should vary from system to system and from move to move within a system, computation of the number of atom-pair energy calculations skipped as a function of τ is trivial and awareness of such a relationship is necessary in determining the values of τ to be considered for use in a given system.

Unfortunately, understanding of the effect of τ on number of atom-pair energy calculations skipped per move gives insight into little else, including real-world impact on computation time or loss of accuracy. To assess these factors, a battery of simulations was run for the C4 monochloro test perturbation of 5-benzyl-*N*-phenyl-1,3,4-oxadiazol-2-amine (**1**, Figure 1.4) unbound and bound to HIV-RT. Of primary interest was monitoring performance to identify bottlenecks in the code; this also proved valuable in determining which methods were most efficient in cases where several methods were possible. Figure 1.21 offers a breakdown of time spent in principal subroutines within the FEP and GB/SA code. Statistics were acquired using optimized code over 5,000 moves of FEP for the C4 monochloro perturbation of **1** bound to HIV-RT with GB/SA solvation and were analyzed with GNU `gprof`. Values of τ were zero and 0.1 Å for the fully rigorous and approximated simulations, respectively.

In Figure 1.21, the top chart represents a fully rigorous simulation where no approximation to the Born energy is applied, $\tau = 0$ Å. Section (a), resected, represents percent time spent in the `CALCEGB` subroutine, which is responsible for the majority of the GB/SA atom-pair energy calculations, as described previously. Additionally, section (b) represents percent time spent in the `EXPJ` subroutine, which is an internal subroutine called by `CALCEGB` for the purposes of computing the exponential function in eq 1.4 for determining atom-pair energies. Thus, in these benchmarks, sections (a) and (b) combined represented nearly 80% of the total simulation time, owing to rigorous atom-pair energy calculations in `CALCEGB`. It is worth noting that the approximation illustrated in Figure 1.19 and the effect modulated by τ in Figure 1.20 apply exclusively to these two subroutines. Subroutine `CALC`, section (c), is responsible for solvent-accessible surface area calculations (the other half of GB/SA theory) and represented nearly 40% of the remaining simulation time (12% of total). Subroutine `READ`, section (d), is primarily responsible for program initialization and disk I/O (a notoriously slow aspect of simulations of large systems), and `GBSASETUP`, section (e), is responsible for initialization of GB/SA energy components

and calculation of 1,2 and 1,3 electrostatic contribution, which is performed once per simulation or sub-block thereof, as previously described. Section (f) represents time spent in all other subroutines combined.

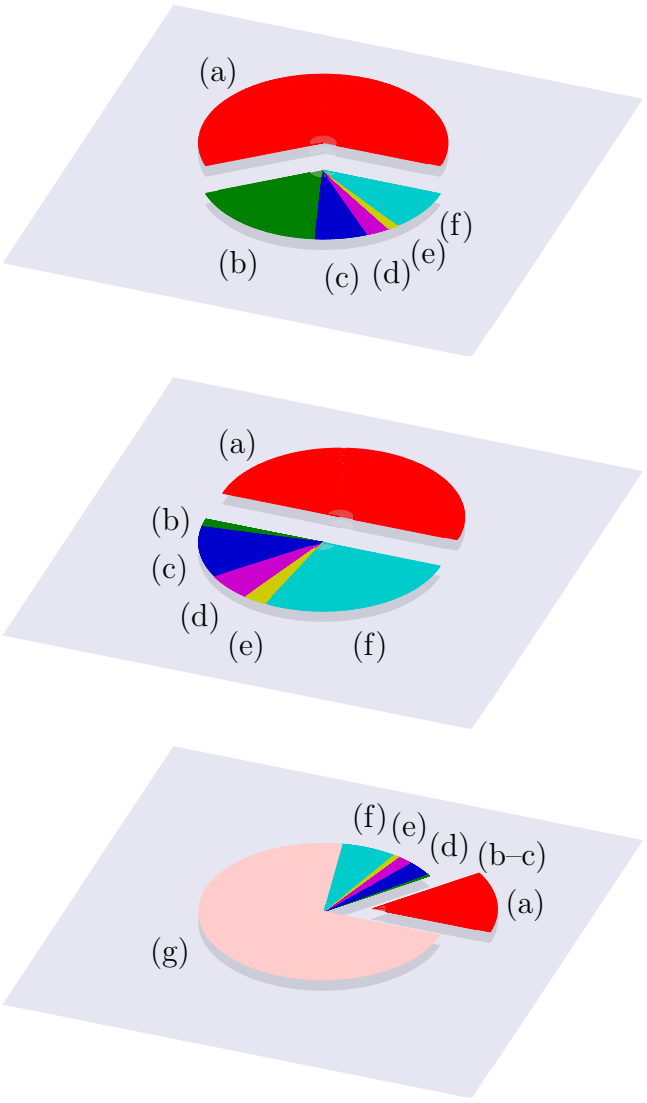


Figure 1.21: GNU `gprof` performance statistics for principal subroutines within the FEP and GB/SA code, where $\tau = 0.0$ Å (top) and $\tau = 0.1$ Å (middle), shown relative to each other at bottom, including time saved. (a) = `CALCEGB`, (b) = `EXPJ`, (c) = `CALC`, (d) = `READ`, (e) = `GBSASETUP`, (f) = All others, (g) = Time saved.

Given this remarkably lopsided distribution, it was gratifying to observe the effect of the approximation in the middle and bottom charts in Figure 1.21. The middle chart represents percent time spent in each principle subroutine in an approximated simulation where a fairly large threshold was applied, $\tau = 0.1$ Å, corresponding to an omission of roughly 99% of atom-pair energy calculations per move according to Figure 1.20. While section (a), again resected and representing percent time spent in the `CALCEGB` subroutine, still accounted for 50% of the total simulation time, notably absent was subroutine `EXPJ`, section (b), whose computational cost diminished from 20% of total simulation time for $\tau = 0$ Å to less than 2% for $\tau = 0.1$ Å. This 90% decrease in exponential function usage corresponds roughly to the omission of roughly 99% of atom-pair energy calculations per move. For comparison, the bottom chart represents subroutine usage in the same approximated simulation relative to total simulation time of the unapproximated, fully rigorous simulation. As illustrated by section (g), this fairly short test simulation of 5,000 moves with a threshold of 0.1 Å required only 28% of the computational time required by that of its unapproximated counterpart. Given the aforementioned caveat that the approximation is not applied until a move is first accepted and therefore that in short simulations a larger percent of total moves are performed in an unapproximated manner, these results were deemed particularly promising with the prospect of applying τ in longer, more realistic simulations. To elucidate this suspected trend, additional GB/SA FEP simulations were run for the C4 monochloro perturbation of **1** bound to HIV-RT with $\tau = 0$, 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4} Å for 100, 1,000, 2,500, and 5,000 moves. For statistical purposes, each combination of simulation length and threshold was repeated along ten unique Monte Carlo trajectories by utilizing different seeds for the pseudo-random number generator. Relative simulation times, averaged over the ten trajectories, are plotted for each simulation length as a function of threshold in Figure 1.22.

Results in Figure 1.22 confirmed the suspicion that the effect of τ on simulation time would be more dramatic with increasing simulation length; however, the trend was

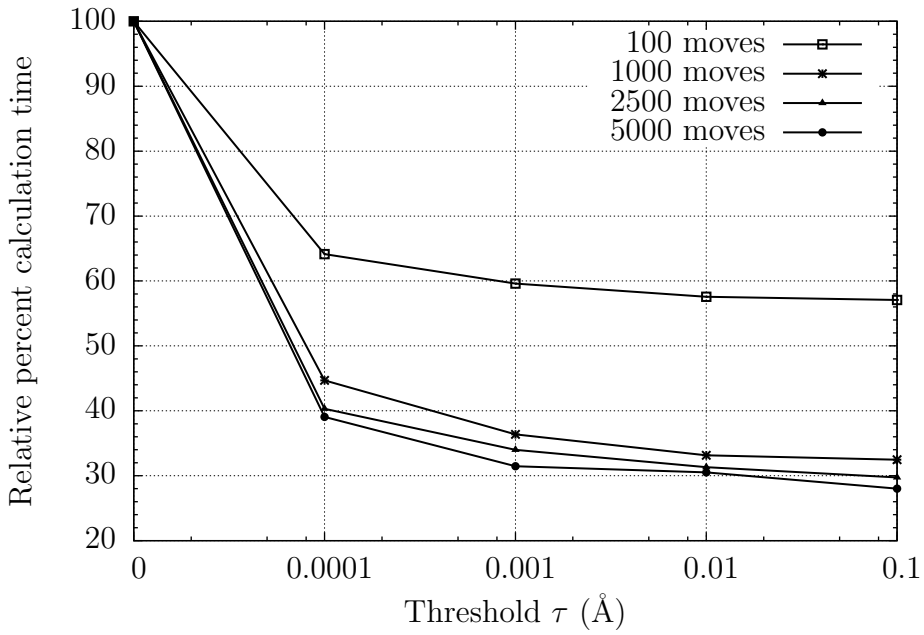


Figure 1.22: Relative simulation time as a function of threshold for the C4 monochloro perturbation of **1** bound to HIV-RT. Times were averaged over ten unique trajectories. Standard deviations ranged from $\sigma = 10^{-4}$ for 100-move simulations to $\sigma = 10^{-2}$ for 5,000-move simulations; all trends were statistically significant.

found to be less remarkable than anticipated: simulations consisting of 1,000, 2,500, and 5,000 moves all experienced nearly the same amount of speed-up at each value of τ. Simulations consisting of 100 moves were afforded little benefit in terms of speed-up from the approximation, owing to the aforementioned caveat. Extrapolation of the data might lead one to conclude that for a simulation consisting of 1.0×10^6 moves, the maximum speed-up one might see would be 75% at any threshold greater than 10^{-3} Å. To that end, increasing the threshold orders of magnitude beyond 10^{-3} Å afforded little benefit in terms of simulation time, since by $\tau = 10^{-3}$ Å the vast majority of atom pairs were already being skipped each move. Given that any increase in threshold allows for the possibility of additional error, for simulations consisting of millions of moves, values between $\tau = 10^{-3}$ Å and $\tau = 10^{-2}$ Å appear to be optimal.

On the other side of the argument on the validity of our approximation lies accuracy; regardless of how fast one might be able to compute relative free energies of binding, those

values are meaningless if they aren't correct. Thus, having elucidated the impact of τ on computational efficiency and with an understanding of the effect of simulation length on τ, it was of great interest to better understand the effect of our approximation on simulation accuracy and the statistical accumulation of error. Population of a Monte Carlo ensemble is a cumulative process, with the acceptance of energies (and therefore properties) of any given move being either directly or indirectly related to those of previous moves. Therefore, when applying an approximation such as ours, one must be cognizant of the fact that approximated energies are accepted or rejected based on other approximated energies. Unchecked, the accumulation of error over the course of millions of moves can become substantial as the trajectory of approximated energies "drifts" from that of the true (unapproximated) energies. To investigate this, free energies from the set of GB/SA FEP simulations on the C4 monochloro perturbation of **1** bound to HIV-RT with $\tau = 0$, 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4} Å for 100, 1,000, 2,500, and 5,000 moves were evaluated. Results from the unapproximated ($\tau = 0$ Å) series were taken as the reference, and results from the approximated series ($\tau = 10^{-1}$, 10^{-2}, 10^{-3}, and 10^{-4} Å) were used to elucidate accumulation of error as a function of threshold and/or number of moves. Results, averaged over ten unique trajectories, were statistically significant ($\sigma < 10^{-2}$) and are shown in Figure 1.23.

The effect of threshold on the accumulation of error was remarkable, primarily in its modesty: the largest threshold considered ($\tau = 0.1$ Å) afforded accumulation of only 0.028% error relative to $\tau = 0$ Å over the course of 5,000 moves; smaller thresholds afforded accumulation of even less, from 0.012% error ($\tau = 0.01$ Å) to 0.0002% error ($\tau = 0.0001$ Å) over 5,000 moves. Also evident in Figure 1.23 is the unsurprising trend that drift accumulates most quickly at larger values of τ; accumulated error with $\tau = 0.0001$ Å showed essentially no dependence on number of moves, whereas it showed a much greater dependence at $\tau = 0.01$ Å and $\tau = 0.1$ Å. Notably, drift at $\tau = 0.001$ Å remained remarkably stable, which when coupled with results in Figure 1.22 makes it

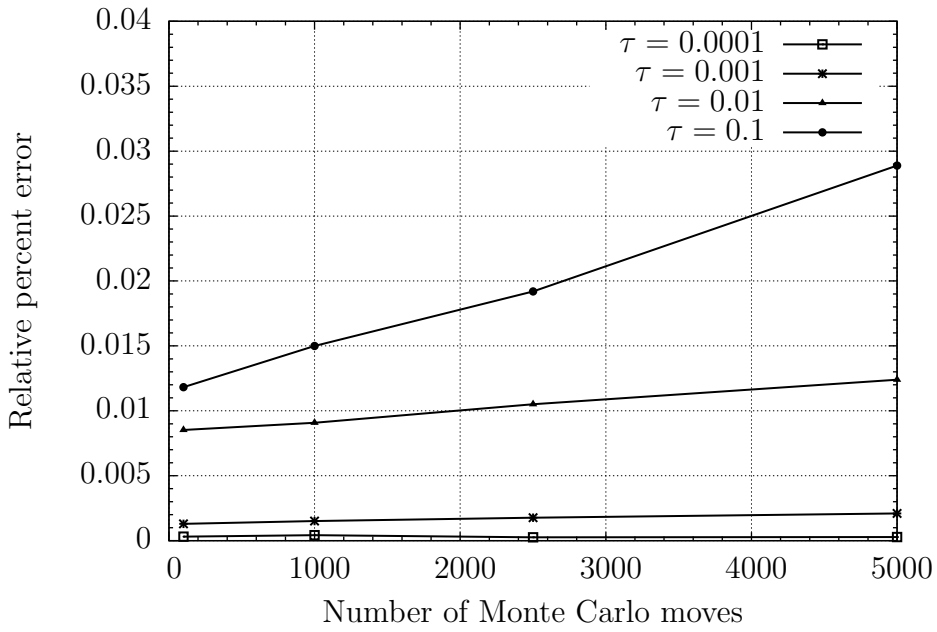


Figure 1.23: Accumulation of error (drift) as a function of MC moves and threshold for the C4 monochloro perturbation of **1** bound to HIV-RT.

a very attractive choice as an optimal threshold value, offering speed-up comparable to larger thresholds without sacrificing susceptibility to drift.

It occurred to us that to combat accumulation of error in FEP simulations using approximations such as ours, unapproximated energy calculations could be performed periodically throughout the course of a simulation, where saved atom-pair Born energies would be erased and new reference energies computed and stored for all atom pairs. In doing do, any drift that had accumulated since the previous reference calculation would be reset, and subsequent approximated energies would computed based on the new reference set rather than an old and substantially approximated one. In most simulations, each window is divided into blocks to simplify the collection and analysis of statistical data for that window. In such cases, the system is reinitialized at the start of each new block, in effect zeroing any drift accumulated during the previous block and keeping total error low. However, restructuring a simulation by dividing it into more blocks consisting of fewer moves per block in order to further minimize drift by reinitializing the system more

frequently is a tedious and largely unnecessary procedure. Thus, a "reset" parameter Ω_τ was implemented such that unapproximated energy calculations would be performed every Ω_τ moves within a block. Such a modification to the algorithm was trivial and was expected to lower the total error of the ensemble over the course of millions of moves.

Unfortunately, not only does the inclusion of Ω_τ necessitate that a full energy calculation be performed for all atom pairs once for each period, it also necessitates that all energy calculations thereafter be performed in an unapproximated manner until the next move is accepted. Thus, low values of Ω_τ negate τ, and so optimal values of Ω_τ could be determined alongside τ for any given system. We conservatively estimate that for full-length simulations of large systems where more than 30,000 moves are made per block, $\Omega_\tau = 5,000$ should be sufficient in keeping drift low while maintaining computational efficiency.

Content with the computational performance of our approximation and pleased with its remarkable tolerance against accumulating error in short simulations, it was desirable to further benchmark its performance in longer simulations. To begin, for the C4 monochloro test perturbation of **1** in the unbound form, full-length FEP simulations were run at $\tau = 10^{-1}$, 10^{-2}, 10^{-3}, and 10^{-4} Å for 1.66×10^5 configurations of equilibration followed by 5.0×10^5 configurations of averaging for each of 14 windows of doublewide sampling; these values correspond to 1/60 of those used in simulations with TIP4P explicit water where solute moves are attempted once every 60 moves. For statistical purposes, each simulation was repeated along five unique Monte Carlo trajectories by utilizing different seeds for the pseudo-random number generator. Averages of the five trajectories and their standard deviations at each threshold are plotted in Figure 1.24.

The data in Figure 1.24 indicate that over the course of five unique trajectories, the statistical impact of τ on $\Delta G_{\text{unbound}}(\mathbf{A} \longrightarrow \mathbf{B})$ was insignificant for long runs; simulations with a large value of τ were just as likely to yield reference free energies as simulations with a small value of τ. Intrigued, the analogous simulations were then performed for

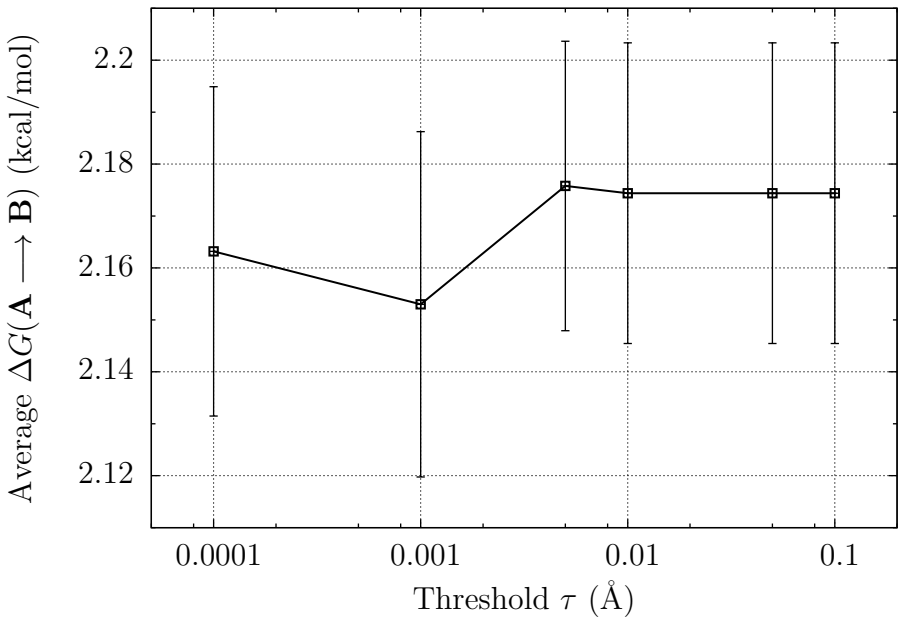


Figure 1.24: Averages of five trajectories at each threshold and their standard deviations for the unbound C4 monochloro perturbation of **1**.

the C4 monochloro test perturbation of **1** bound to HIV-RT, and averages of the five trajectories and their standard deviations at each threshold are plotted in Figure 1.25.

The data in Figure 1.25 indicate again that over the course of five unique trajectories, the statistical impact of τ on $\Delta G_{\text{bound}}(\mathbf{AP} \longrightarrow \mathbf{BP})$ was insignificant for long runs; simulations with a large value of τ were just as likely to yield reference free energies as simulations with a small value of τ. Fluctuations among trajectories at the same τ were notably larger for the simulations in the bound form (σ = 0.1–0.4) than they were in the unbound form (σ = 0.0028–0.033), which is expected owing simply to the difference in size between the unbound (30 atoms) and bound (2,728 atoms) systems. Combining the data in Figures 1.24 and 1.25, free energies of binding at the different thresholds were determined along with their statistical significance, as shown in Figure 1.26.

Gratifyingly, the same statistical insignificance of the effect of τ on the unbound and bound systems carried over to the determination of relative free energies of binding; for large systems simulated over many Monte Carlo moves, the amount of error introduced

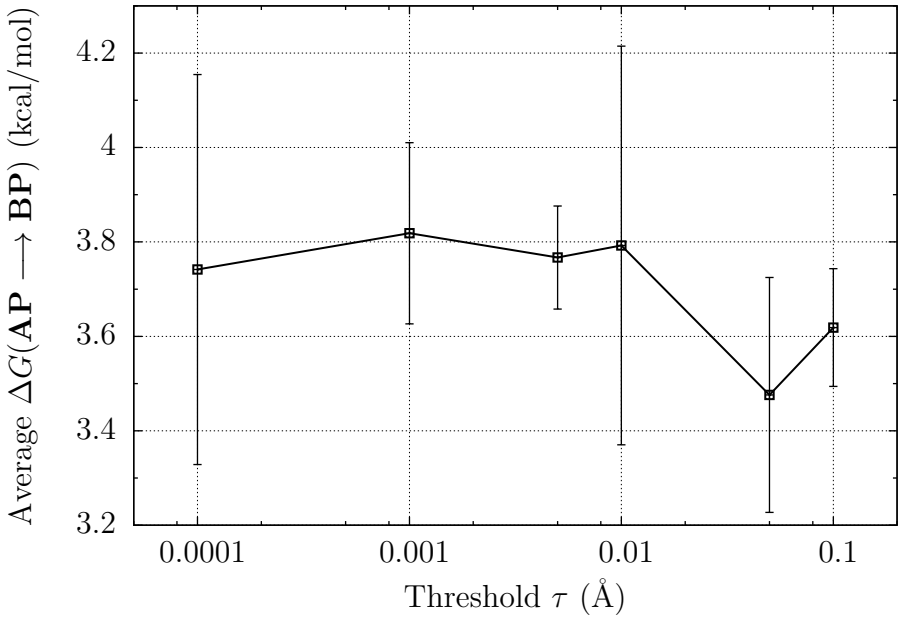


Figure 1.25: Averages of five trajectories at each threshold and their standard deviations for the bound C4 monochloro perturbation of **1**.

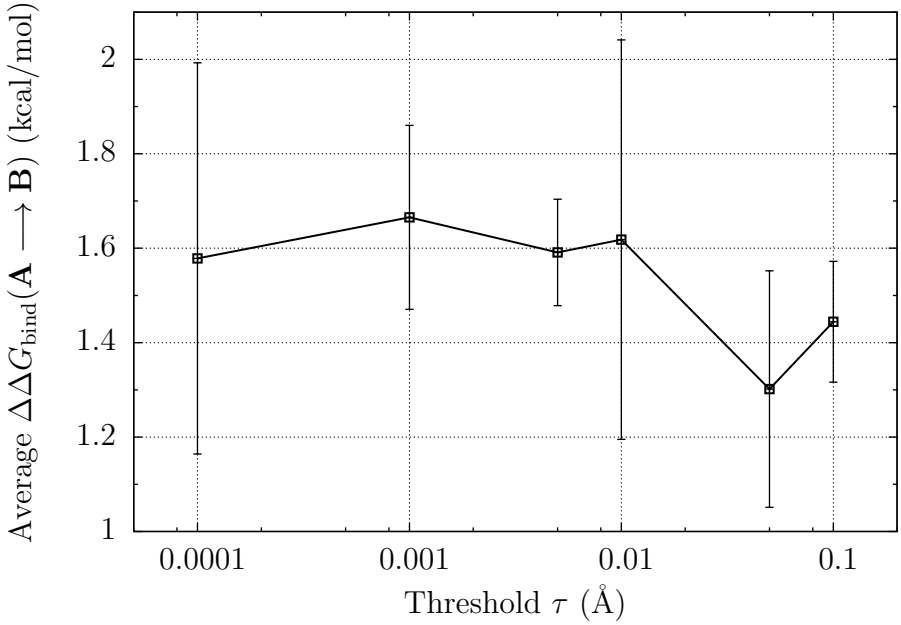


Figure 1.26: Averages of five trajectories at each threshold and their standard deviations for the free energy of binding of the C4 monochloro perturbation of **1** bound to HIV-RT.

into the ensemble by the approximation was found to be less than the amount of statistical noise generated by the Metropolis Monte Carlo sampling procedure itself. Thus, to a degree, these data suggest that any threshold could be utilized with a certain amount of assurance that the results obtained would be within the statistically tolerated margin of error of the unapproximated, fully rigorous simulation.

1.2.6 Chlorine scan and comparison to TIP4P.

In the search for non-nucleoside inhibitors of HIV-1 reverse transcriptase, compounds consisting of a 1,3,4-oxadiazole core linking a phenyl and an anilinyl ring such as **1** have proven to be promising leads. One of the initial steps in optimizing such a ligand is often to perform a substituent "scan" on each of the benzyl and anilinyl rings, wherein a substituent such as a chlorine is perturbed to a hydrogen on each of the ring carbons, as illustrated along with our labeling convention in Figure 1.4. By performing FEP calculations and determining the relative free energies of binding afforded by placing the substituent at various positions on the core, one can elucidate which locations of the given substitution play the most important role in stabilizing host–guest interactions. In the development of this ligand, a chlorine scan was performed originally with TIP4P explicit water,[59] and further development based on these studies has led to several promising drug candidates.[60] It was therefore of great interest to see how well free energies of binding could be attained via a chlorine scan using GB/SA, and how similarly the results compared with those from the earlier TIP4P simulations.

Supplementing the calculations from Figure 1.26, full-length FEP simulations were performed for each bound and unbound monochloro perturbation of **1**, with no threshold and with the optimized values $\tau = 10^{-3}$ Å and $\Omega_\tau = 5{,}000$, for 1.66×10^5 configurations of equilibration followed by 5.0×10^5 configurations of averaging for each of 14 windows of doublewide sampling. The results of the chlorine scan with unapproximated GB/SA are shown in Table 1.4 and the results with approximated GB/SA are shown in Table 1.5.

Table 1.4: Results of the chlorine scan of **1** with unapproximated GB/SA. Values are in kcal/mol.

Structure	ΔG_{bound}	$\Delta G_{\text{unbound}}$	$\Delta\Delta G_{\text{bind}}$
C2	3.887	2.131	1.756
C2′	7.491	1.347	6.144
C3	1.717	−1.158	2.875
C3′	−0.400	0.907	−1.307
C4	5.650	3.276	2.374
C4′	0.304	2.215	−1.911
C5	1.048	−0.886	1.934
C5′	2.522	1.370	1.152
C6	−5.502	−2.919	−2.583
C6′	6.838	1.605	5.233

Table 1.5: Results of the chlorine scan of **1** with approximated GB/SA with $\tau = 10^{-3}$ Å and $\Omega_\tau = 5{,}000$. Values are in kcal/mol.

Structure	ΔG_{bound}	$\Delta G_{\text{unbound}}$	$\Delta\Delta G_{\text{bind}}$
C2	3.899	2.131	1.768
C2′	6.872	1.347	5.525
C3	1.725	−1.158	2.883
C3′	−0.333	0.907	−1.24
C4	5.648	3.276	2.372
C4	0.101	2.174	−2.073
C5	0.943	−0.886	1.829
C5′	2.349	1.370	0.979
C6	−5.714	−2.919	−2.795
C6′	6.708	1.605	5.103

In Tables 1.4 and 1.5, positive values indicate favorable change in free energy of binding for the chlorine substitution (Cl $\longrightarrow$ H). Results further suggest a minimal effect of our approximation on the accuracy of computed free energies; unsigned errors between approximated and rigorous results ranged from as low as 0.002 kcal/mol for the C4 perturbation to only 0.2 kcal/mol for the C6 perturbation, for a mean unsigned error of 0.15 kcal/mol for the series. Simulation times in the bound state were 6 weeks for each rigorous calculation and 10 days for each approximated calculation when windows were run serially; dividing windows into parallel jobs reduced simulation times to 24 hours for each approximated GB/SA FEP calculation. Computational demand was negligible for

simulations in the unbound state.

Results of the previous chlorine scan with TIP4P are shown in Table 1.6, taken from reference 60, and a graphical comparison of relative free energies of binding from Tables 1.4, 1.5, and 1.6 is presented in Figure 1.27. Again, positive values indicate favorable change in free energy of binding. Values are in kcal/mol.

Table 1.6: Results of the chlorine scan of **1** with TIP4P, taken from reference 60. Values are in kcal/mol.

Structure	ΔG_{bound}	$\Delta G_{\text{unbound}}$	$\Delta\Delta G_{\text{bind}}$
C2	2.713	0.505	2.208
C2′	4.863	1.431	3.432
C3	1.327	−2.000	3.327
C3′	−4.013	0.691	−4.704
C4	6.618	2.415	4.203
C4′	1.010	1.806	−0.796
C5	−0.542	−1.707	1.165
C5′	1.714	0.801	0.913
C6	−6.863	−4.264	−2.599
C6′	4.887	1.115	3.772

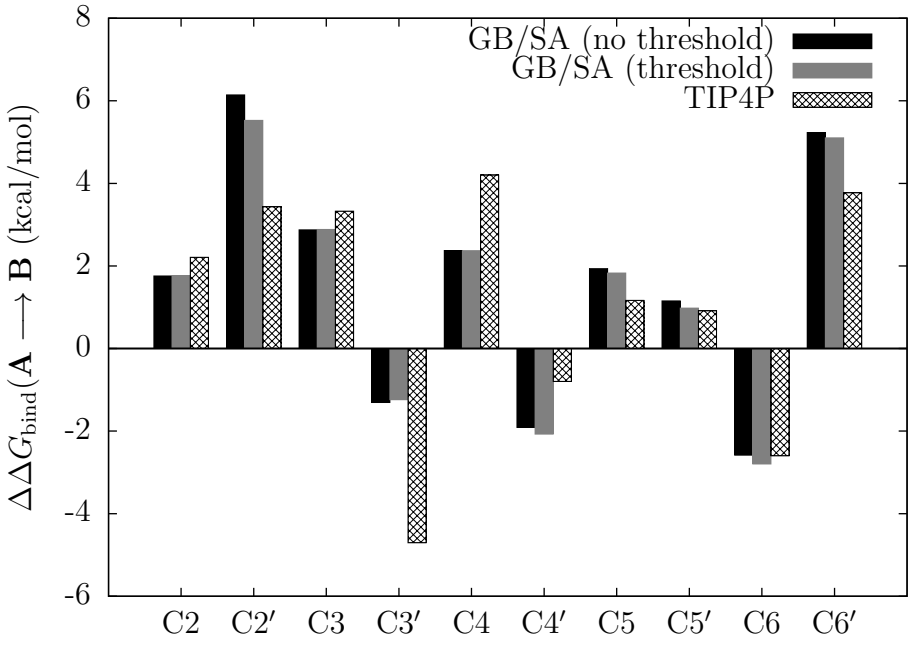


Figure 1.27: Chlorine scan of free energies of binding comparing GB/SA and TIP4P.

Figure 1.27 illustrates qualitative-to-quantitative agreement between free energies of binding computed using GB/SA and TIP4P solvation. For C2, C3, and C5′ derivatives, unsigned error was less than 0.5 kcal/mol for simulations with GB/SA solvation relative to simulations with TIP4P explicit water. Surprisingly, mean unsigned error for GB/SA relative to TIP4P was lower for the approximated series (1.18 kcal/mol) than for the unapproximated series (1.24 kcal/mol); simulations with unapproximated GB/SA solvation tended to overestimate free energies of binding moreso than did simulations with approximated GB/SA solvation relative to TIP4P. A reasonable explanation for the tendency for GB/SA to overestimate free energies in FEP can be found in eq 1.6, which defines ΔG as a statistical function of energy differences E; when GB/SA is used, values of E are instead substituted by G, so contribution to the statistical approximation of ΔG would theoretically be expected to be over-counted. Nevertheless, based on the GB/SA chlorine scan results, substitution with chlorine at C2′ (5.5 kcal/mol) and C6′ (5.1 kcal/mol) was predicted to enhance free energy of binding most dramatically, followed by substitution at C3 (2.8 kcal/mol) and C4 (2.3 kcal/mol), whereas the TIP4P scan results predicted substitution by chlorine at the C4 position (4.2 kcal/mol) to enhance free energy of binding most dramatically, followed by substitution at C2′ (3.4 kcal/mol) and C6′ (1.3 kcal/mol). Notably, optimization of the lead was pursued previously[60] based on the TIP4P scan results in 2010[59] and the structure was elaborated upon significantly at the C4 position, along with fluorine substitution at the C2′ and C6′ positions for conformational purposes. The resulting structure, Figure 1.28, was found to be highly active in preliminary anti-retroviral assays. It was found[60] that the C4 elaboration afforded favorable polar contacts with a solvent-accessible region and that the C2′ and C6′ substitutions afforded conformational pre-arrangement of the benzyl ring, lowering the energetic penalty of conformational focusing. The GB/SA scan results corroborate this substitution pattern, suggesting that GB/SA can could be used as a viable solvation model in applications in drug design.

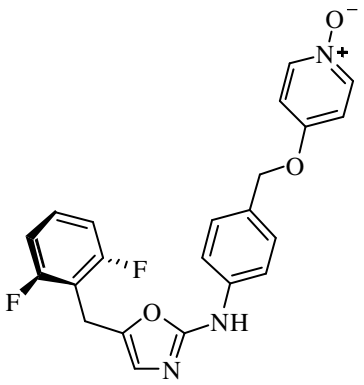


Figure 1.28: FEP-optimized structure corroborating GB/SA chlorine scan results.

1.3 Summary and conclusions.

A description of our implementation of GB/SA with FEP was presented, including an approximation to calculating the total Born energy of the system. Early benchmarks demonstrated that existing GB/SA methodologies were insufficient for the purposes of calculating free energies of binding, and FEP simulations with GB/SA solvation were too computationally expensive to be used with any practicality. We have shown that with our approximation, τ and Ω_τ improved efficiency and minimized error, and both can be fine-tuned to accommodate a given system or investigator's needs. For our test system, the influence of τ on accuracy of the free energies of binding was negligible, with any error introduced by the approximation falling well below the statistical error of the Metropolis Monte Carlo algorithm. Moreover, speed-up of up to $3.8x$ was observed at $\tau > 0.005$ Å, with $3x$ speed-up observed even at τ as low as $\tau = 10^{-4}$ Å, making GB/SA a viable solvent choice for FEP of large systems. In a substituent scan on our test system, agreement between GB/SA and TIP4P was quantitative to qualitative, and simulations with GB/SA solvation suggested the same substitution pattern found to give high anti-retroviral activity as predicted by previous simulations with TIP4P explicit water.

Chapter 2

E/Z energetics for molecular modeling and design.

2.1 Introduction.

Knowledge of the conformational energetics of small molecules is essential in many areas of chemistry, including organic synthesis and molecular design.[61] The conformational preferences for small molecules are well known to carry over to macromolecular structures; for example, the ca. 3 kcal/mol preference for the Z conformer of N-methylacetamide relative to the E alternative is primarily responsible for the rarity of *cis*-peptide bonds in proteins.[62] In the context of molecular modeling and drug design, molecules where rotation about a single bond leads to E and Z conformers that are energetically well separated by an intervening potential-energy barrier are of considerable interest. Besides amides, molecules in this category include other derivatives of carboxylic acids, aldehydes, or ketones such as esters, carbamates, carbonates, ureas, amidines, hydrazones, and oximes. The importance of these functional groups is enhanced by their common occurrence in combinatorial libraries, commercial screening collections, and in molecules of pharmacological interest. Though there have been prior computational studies of molecules featuring E/Z equilibria, most studies have focused on one or two functional groups using Hartree–Fock (HF), B3LYP-based density functional, or second-order Møller–Plesset (MP2) theory.[63–83] Some classic studies include those of Wiberg and co-workers on formic acid, acetic acid, methyl formate, and methyl acetate.[63] N-methylacetamide has also received much attention owing to its status as a model for the peptide bond.[64,73,76,77,83] The need for quantum

mechanical investigations in this area is magnified by the fact that experimental studies of E/Z equilibria are often challenging owing to a very small population of the higher-energy conformer.

2.1.1 Investigative focus.

In seeking enzyme inhibitors through de novo design or virtual screening,[84,85] questions often arise about the likelihood of E and Z conformers. For example, in docking studies, one is regularly confronted with computed structures for complexes, or "poses," such as in Figure 2.1, left, where the ligand features an E or Z conformation. The scoring with docking software is still improving and such poses may score well,[86] though the E conformer for the ester in this case is unreasonable.[61] Alternately, one may be confronted with a crystal structure, such as in Figure 2.1, right, where the thiourea moiety is in an E,Z configuration.[87] If one suspected there to be an associated energetic penalty, alternative designs might be pursued to achieve enhanced potency. Given many such examples, it was desirable to pursue clarification of conformational energetics through reliable quantum mechanical calculations on prototypical molecules featuring E and Z

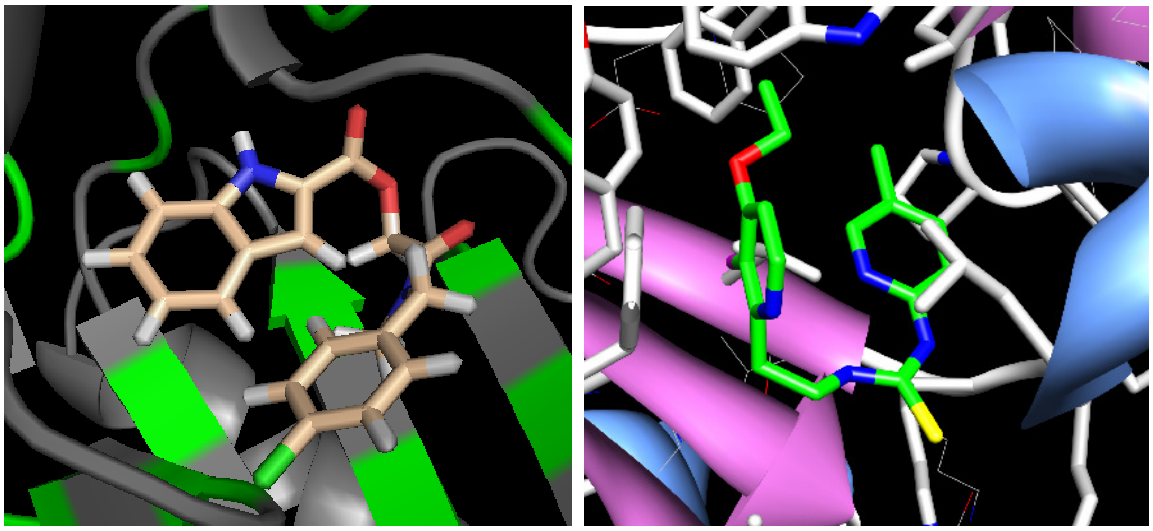


Figure 2.1: Structure of an ester-containing molecule docked into HIV-1 reverse transcriptase (RT), left, and the 1dtt crystal structure of an analog of trovirdine bound to HIV-RT illustrating an E,Z conformer for a thiourea moiety, right.

conformers. The findings would be valuable as a basis for the improvement of scoring functions for docking software,[88] the refinement of crystal structures, and development of molecular mechanics force fields for use in modeling organic and biomolecular systems.[89,90] Thus, for broader coverage of E/Z equilibria at a higher and consistent level of theory, the 18 pairs of conformers illustrated in Figures 2.2, 2.3, and 2.4 were examined using composite ab initio methods.

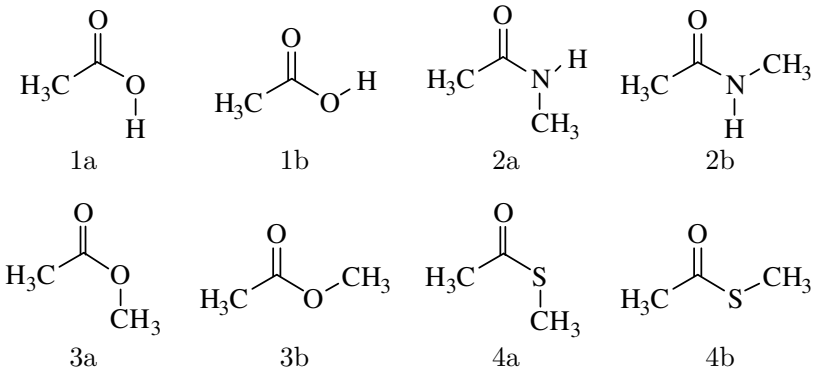


Figure 2.2: Molecules in the RCOX set.

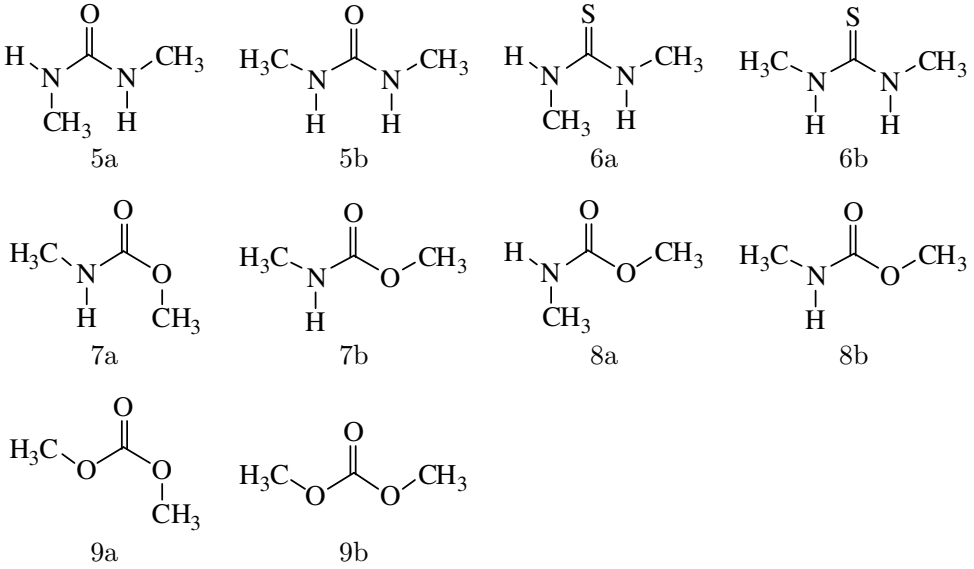


Figure 2.3: Molecules in the RXCOYR set.

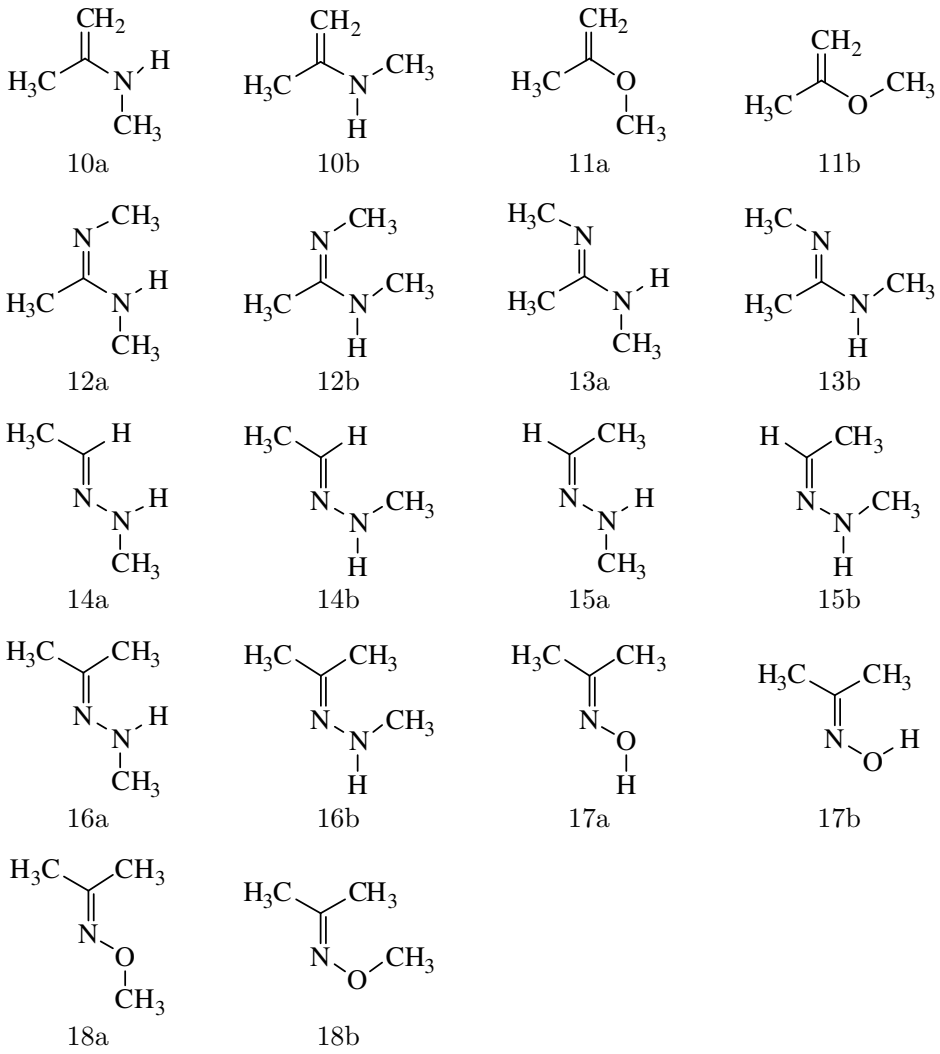


Figure 2.4: Molecules in the C=C&N set.

2.1.2 Computational details.

All ab initio and DFT calculations were carried out using the Gaussian03 program.[91] The G3 and G3B3 methods were applied to compute structures, dipole moments, vibrational frequencies, energies at 0 K, and enthalpies and free energies at 298 K.[14,92] In the G3 method, the initial geometry optimization and vibrational frequency and zero-point energy calculations are performed at the 6-31G(d) level. The geometry is then refined including electron correlation at the MP2(full)/6-31G(d) level. A series of single-point energy calculations follows, using MP2/G3large (a basis set with core correlation), MP4/6-

31G(d), and QCISD(T)/6-31G(d), with spin-orbit and other higher corrections. The G3B3 approach particularly improves the initial geometry, vibrational frequencies and zero-point energy by starting with a B3LYP/6-31G(d) geometry optimization. Estimates of free energies of hydration were made for all conformers using the generalized Born / surface area (GB/SA) approach,[27,28] as implemented in *BOSS*.[33,34] Structures were optimized using the OPLS/CM1A force field,[90] and the GB/SA calculations were performed with CM1A atomic charges scaled by 1.07.[34]

2.2 Results and discussion.

2.2.1 E/Z conformers.

The 18 pairs of conformers that were investigated are shown in Figures 2.2, 2.3, and 2.4. The RCOX set consisted of a prototypical carboxylic acid, secondary amide, ester, and thioester. The RXCOYR set contained a urea, thiourea, carbamate (urethane), and carbonate, while the C=C&N set covered an enamine, an enol ether, amidines, hydrazones, and oximes. For all pairs, conformer **a** was the E conformer and conformer **b** was the Z conformer. For amine derivatives, secondary cases $RNHCH_3$ were considered, and although E and Z are well defined for tertiary cases $RNR'R''$, the E/Z preferences for them are generally well predicted by simple steric considerations. It should be noted that conformers **7b** and **8b** were the same, which simplifies the presentation of results. For the RCOX set, both G3 and G3B3 calculations were performed, while the RXCOYR and C=C&N sets were investigated only using the G3B3 method.

2.2.1.1 Results for the RCOX set.

For conformer pairs **1**–**4**, the G3 and G3B3 results are given in Table 2.1. In all cases, the relative values are given for conformer **a** minus conformer **b** ($E - Z$). The G3 and G3B3 energetic results generally agreed to within 0.1 kcal/mol, and thermal corrections to the vibrational energies were nearly the same between conformers, so there was little

difference between the results for ΔE (0 K) and ΔH (298 K). The computed entropy changes were generally small, though there was some sensitivity to the treatment of low-frequency vibrations.

Table 2.1: Computed differences in energies (kcal/mol) and dipole moments (D) from G3 and G3B3 calculations for the RCOX set.

Pair	ΔE (0 K)	ΔH (298 K)	ΔG (298 K)	$\Delta\mu$
		G3		
1	5.08	5.11	5.15	2.93
2	2.42	2.22	3.11	0.32
3	7.48	7.46	7.47	3.10
4	4.63	4.43	4.60	3.15
		G3B3		
1	5.11	5.11	5.27	2.92
2	2.34	2.26	2.67	0.32
3	7.42	7.41	7.42	3.10
4	4.63	4.41	4.38	3.19

For acetic acid (pair **1**), the *E* conformer was found to be 5.1 kcal/mol higher in energy than the *Z* conformer from the G3 and G3B3 calculations. This is in accord with an MP4/cc-pVTZ result of 5.38 kcal/mol[78] while lower levels of theory have generally given larger differences.[63] An experimental result was not available for comparison, though the *E* conformer has been detected in an argon matrix at 8 K[93] and the best estimate for the energy difference for formic acid is about 1 kcal/mol smaller at 4.21 kcal/mol.[75] For *N*-ethylacetamide (pair **2**), the present results concur with other high-level calculations and experiments that found the enthalpy difference at 298 K to be in the 2.1–2.5 kcal/mol range.[73,76,77,83,85] The difference was found to diminish to 1.0–1.2 kcal/mol for *N*-methylformamide owing to reduced steric crowding in the *E* form.[73,94]

Similarly, the G3 and G3B3 results for methyl acetate (pair **3**) were in line with the energy difference of 7.72 kcal/mol from LMP2/cc-pVTZ(-f) calculations,[73] while again older values have been somewhat higher.[63,68,69] For methyl formate, the LMP2 energy difference was found to be only 5.35 kcal/mol.[80] Besides the steric effects favoring *Z*, the

E conformer of carboxylic acids and esters is also known to be destabilized by unfavorable dipole–dipole interactions or lone pair–lone pair repulsion between the oxygen atoms.[70,71] As indicated in Table 2.1, the dipole moments for *E* acids and esters were found to be ca. 3 D higher than for the *Z* forms. Overall, the population of *E* carboxylic esters is generally known to be vanishingly low and so invoking acyclic esters in this conformation is improper.[95,96] The $E - Z$ energy difference for the corresponding thioester (pair **4**) moderated to 4.6 kcal/mol owing in part to the longer C–S than C–O bonds, which has the effect of lowering the 1,4-CC steric penalty for the *E* conformer. Again, the difference would be expected to be less for methyl thioformate, an assumption that has been confirmed by NMR studies indicating a free energy difference of ca. 1.3 kcal/mol.[94]

2.2.1.2 Results for the RXCOYR and C=C&N sets.

The G3B3 results for the remaining pairs are given in Table 2.2. The results are largely understandable in terms of the strong preference for the *Z* conformers in pairs **1**–**3** and additional steric and electronic effects, as described. Interestingly, in comparison to *N*-methylacetamide, the *Z,Z* over *E,Z* energetic preference for 1,3-dimethylurea (pair **5**) was found to be reduced to 1.06 kcal/mol, and the *E,Z* conformer of 1,3-dimethylthiourea (conformer **6a**) was found to be favored by 0.17 kcal/mol. In prior work, MP2/aug-cc-pVDZ results favored the *Z* conformer of methylurea and methylthiourea by 1.25 and 0.70 kcal/mol,[81,82] and MP2/6-31G(d) results have indicated a preference for the *Z,Z* conformation over the *E,Z* conformation for 1,3-dimethylurea by 1.72 kcal/mol.[72] The G3B3 result for pair **5** is expected to be more accurate and indicates that there would only be a small intrinsic penalty for incorporating an *E,Z*-urea substructure in a molecular design.

Moreover, the *E,Z*-thiourea fragment like in Figure 2.1 was found to be preferred over the *Z,Z* alternative. In fact, there have been extensive NMR studies of the conformational equilibria for pair **6** in multiple solvents with the conclusion that the *E,Z* conformer is

Table 2.2: Computed differences in energies (kcal/mol) and dipole moments (D) from G3B3 calculations for the RXCOYR and C=C&N sets.

Pair	ΔE (0 K)	ΔH (298 K)	ΔG (298 K)	$\Delta\mu$
		RXCOYR		
5	1.06	1.03	1.09	0.50
6	−0.17	−0.09	−0.55	0.80
7	7.47	7.30	8.18	3.05
8	1.24	1.15	1.75	0.34
9	3.03	2.99	3.09	3.60
		C=C&N		
10	2.67	2.65	2.60	−0.13
11	4.47	4.61	4.01	1.27
12	−4.06	−4.05	−3.95	0.31
13	3.13	3.00	3.44	−0.04
14	−0.16	−0.03	−0.35	−0.01
15	−3.30	−3.21	−3.28	0.37
16	−2.54	−2.43	−2.69	0.54
(17)†	−6.04	−6.00	−6.35	−2.94
(18)†	−19.07	−18.34	−20.40	−2.66

†The planar *Z* form **b** is a transition state.

lower by ca. 1 kcal/mol in free energy than the *Z*,*Z* conformer and that the *E*,*E* form is not populated.[74] Thus, the electronic energy from MP2/cc-pVDZ calculations without zero-point or other corrections in that study appears to lead to the wrong qualitative conclusion by favoring the *Z*,*Z* conformer by 0.38 kcal/mol.[84] In summary, the *E*,*Z* conformer for ureas was found to be relatively more favorable than the *E* conformer of secondary amides, and the *E*,*Z* conformer for the prototypical 1,3-dialkylthiourea (pair **6**) was determined to be the lowest in energy. In view of the small differences in dipole moments for pairs **5** and **6** in Table 2.2, these preferences are not expected to be strongly influenced by medium effects. A possible contributor to the increased favorability of the *E*,*Z* geometry in the ureas is π-electron donation (amide resonance $^{+}$N=C–X^{-}), which increases the partial negative charge on the oxygen or sulfur atom and improves the electrostatic interaction with the *syn*-hydrogen on nitrogen in the *E* substructure.

The results for pairs **7**–**9** in Table 2.2 present an interesting contrast. For 1,3-dimethyl

carbamate (pair **7**), rotation of the methoxy group to the *E* form was found to be similarly unfavorable as for the methyl ester (pair **3**), while rotation of the *N*-methyl group in going from **8b** to **8a** was about 1 kcal/mol less costly than for the methyl amide (pair **2**). The relative G3B3 energies for the three conformers of the carbamate, *Z,Z* (**7b**), *Z,E* (**7a**), and *E,Z* (**8a**) were 0.0, 7.47, and 1.25 kcal/mol, respectively. Thus, as for the dimethyl urea (pair **5**), the penalty for rotation of the *N*-methyl group in the carbamate to the *E* form was not large; however, it appears that an *E* geometry for the ester fragment remains too high in energy for significant population under normal conditions. The possibility for an *E*-ester substructure was found to be significantly improved for dimethyl carbonate (pair **9**), for which the *E,Z* conformer was only 3.03 kcal/mol higher in energy than the *Z,Z* form. The 4–5 kcal/mol diminution relative to pairs **3** or **7** likely stems from destabilization of the *Z,Z* conformer by repulsion between the lone pairs on the methoxy oxygen atoms. Previous MP2/6-31G(d) results for the relative electronic energies of the *Z,Z*, *E,Z*, and *E,E* conformers of pair **9** were 0.0, 3.36, and 26.73 kcal/mol.[71]

Turning to the molecules in Figure 2.4, pairs **10** (*N*-methyl-2-aminopropene) and **11** (2-methoxypropene) are the olefinic analogs of pair **2** and **3**. Accordingly, the energetic preference remained the same, significantly favoring the *Z* conformers by 2.67 kcal/mol (pair **10**) and 4.47 kcal/mol (pair **11**). Thus, the 1,4-CC interaction appears to continue to dominate, while the larger energy difference for the ester in pair **3** over the enol ether in pair **11** can be attributed to the addition of the lone-pair repulsion between the oxygens for the *E* conformer of the ester (**3a**). Based on the results mentioned above for formic acid versus acetic acid derivatives, the $E-Z$ energy differences for the corresponding vinyl analogs of pairs **10** and **11** would be expected to be reduced by ca. 1.0 and 2–3 kcal/mol, respectively. Indeed, MP2/6-31G results have provided an $E-Z$ energy difference of about 2.0 kcal/mol for methyl vinyl ether,[97] and our calculations indicated 1.74 kcal/mol for $\Delta E(0\ \mathrm{K})$ using G3B3. It should be noted that the *E* conformers of pairs **10** and **11** were not planar; the G3B3 results for the H_3C–C–X–CH_3 dihedral angles were 40.8° and

37.1° for conformers **10a** and **11a**. Thus, these conformers may be described as skew. Amine nitrogens were also somewhat paramidalized in all structures; for example, the H_3C–C–N–CH_3 dihedral angle in **10b** was 170.6°; however, the H_3C–C–O–CH_3 dihedral angle in conformer **11b** was 180°. The results for the corresponding dihedral angles for all conformers are listed in Table 2.3.

Table 2.3: G3B3 results for key dihedral angles ϕ (degrees).

Conformer	Angle	ϕ	Conformer	ϕ
1a	CCOH	0.0	**1b**	180.0
2a	CCNC	9.8	**2b**	179.9
3a	CCOC	0.3	**3b**	179.9
4a	CCSC	0.0	**4b**	178.3
5a	NCNC	20.4	**5b**	169.4
6a	NCNC	7.0	**6b**	173.9
7a	NCOC	5.6	**7b**	179.9
8a	OCNC	9.8	**8b**	179.9
9a	OCOC	0.0	**9b**	180.0
10a	CCNC	40.8	**10b**	170.6
11a	CCOC	37.1	**11b**	180.0
12a	NCNC	165.0	**12b**	5.0
13a	NCNC	148.5	**13b**	9.0
14a	CNNC	152.0	**14b**	22.3
15a	CNNC	159.7	**15b**	69.2
16a	CNNC	161.3	**16b**	79.7
17a	CNOH	179.9	**(17b)**†	0.0
18a	CNOC	180.0	**(18b)**†	1.4

†Transition state.

For pairs **12** and **13** in Figure 2.4, the structures represent the four conformers for *N,N'*-dimethylacetamidine. The trans-(Z) conformer **13b** can be argued to be the most analogous to (Z)-*N*-methylacetamide (**2b**) and it was found to be the lowest in energy. The relative energies ΔE (0 K) for the other conformers were 1.07, 3.13, and 5.13 kcal/mol for conformers **12a**, **13a**, and **12b**, respectively, at the G3B3 level. The $E - Z$ energy difference in Table 2.2 for pair **13** was also just a little greater than for pair **2**, possibly reflecting diminished electrostatic attraction for N$\cdots$HN in conformer **13a** than for O$\cdots$HN in conformer **2a**. The 1,5-CC interaction in conformer **12b** was

found to be particularly destabilizing as it is similar to a *syn*-pentane interaction, so this conformer is not competitive. Overall, two low-energy conformers are apparent for the dimethylamidine: **13b** and **12a**.

Similarly, for pairs **14** and **15** in Figure 2.4, the four structures are the conformers for the *N*-methyl hydrazone of acetaldehyde, while conformers **16a** and **16b** are the *E* and *Z* possibilities for the *N*-methyl hydrazone of acetone. For pairs **14** and **15**, the lowest energy conformer was found to be **14a** and the relative energies, $\Delta E(0\ \mathrm{K})$, were 0.16, 0.34, and 3.63 kcal/mol for conformers **14b**, **15a**, and **15b**. Thus, the first three conformers were found to be very close in energy with only conformer **15b** being uncompetitive owing to the 1,5-CC interaction. It is also then easy to predict that *Z* conformer **16b** would be higher in energy than the *E* alternative **16a**; the difference of 2.54 kcal/mol is a little smaller in magnitude than what was calculated for pair **15**. A message from this for molecular design is that the conformational diversity of hydrazones of ketones is much less than for hydrazones of aldehydes.

Finally, the oxime (pair **17**) and *O*-methyl oxime (pair **18**) of acetone were considered. In both cases, the planar *Z* form was found to be a transition state with the G3B3 calculations; only the *E* structures were found to be energy minima. The *Z* transition states were 6.04 kcal/mol (pair **17**) and 19.07 kcal/mol (pair **18**) higher in energy than the *E* conformers. The *Z* structures are presumably destabilized by electrostatic repulsion between the lone-pair electrons on N and O and by the *syn*-pentane-like interaction in **18b**. The status of the *Z* structure for acetoxime appears to be sensitive to the computational level. For example, we found with B3LYP/6-31G(d) energy minimizations and vibrational frequency calculations that the C=N–O–H planar *Z* structure is a shallow energy minimum; it was found to be 6.16 kcal/mol above the *E* conformer, separated from conversion to the *E* form by a barrier of 2.0 kcal/mol at a dihedral angle near 70° and with the hydroxyl hydrogen constrained between two of the hydrogens on the *syn*-methyl group, staggering the C=N bond. However, the G3B3 results indicate that for both the

oxime and *O*-methyl oxime, only the *E* conformational energy well exists. Besides the common occurrence of oximes in screening collections, interest in them also continues as the substrates for Beckmann rearrangements. In this case, the dominance of the *E* conformers is relevant for proposed mechanistic schemes.[98]

2.2.2 Summary of *E*/*Z* results.

A summary of the *E*/*Z* free-energy differences for all conformer pairs is provided in Figure 2.5. A positive difference indicates that the *Z* conformer was found to be favored, and a negative difference indicates that the *E* conformer was found to be favored. Some general rules are evident: For rotation about C–X (X = OR, SR, NHR) single bonds in O=C–X, N=C–X, and C=C–X substructures, *Z* conformers are normally preferred in the absence of significant steric effects that preferentially destabilize the *Z* conformer, especially *syn*-pentane-like interactions. The *Z* preference diminishes in the order X = OR ¿ OH ¿ SR ¿ NHR; it is also diminished for thione derivatives, S=C–X, and through additional conjugation as in ureas. Rotation about the N–N and N–O bonds in hydrazones and oximes favors the *E* conformers, especially when reinforced by a *syn*-pentane-like interaction in the *Z* form.

Concerning dipole moments, there is a general correlation in Tables 2.1 and 2.2 such that the conformer with the larger dipole moment is normally higher in energy than the one with the smaller dipole moment. This is reasonable based on electrostatic considerations and contributes to the general preference for *Z* conformers. The largest differences in dipole moments ($\Delta\mu = \mu_E - \mu_Z$) are 3–4 D and these correspond to cases where the *E* conformer is higher in energy by 3–8 kcal/mol. For the oximes, $\Delta\mu$ is ca. -3 D and consistently the *E* conformers are significantly favored. Of course, steric effects modulate the results such that, for example, $\Delta\mu$ is small for pairs **12**, **15** and **16**, but the *E* conformers are strongly favored owing to the 1,5-CC interactions in the *Z* conformers.

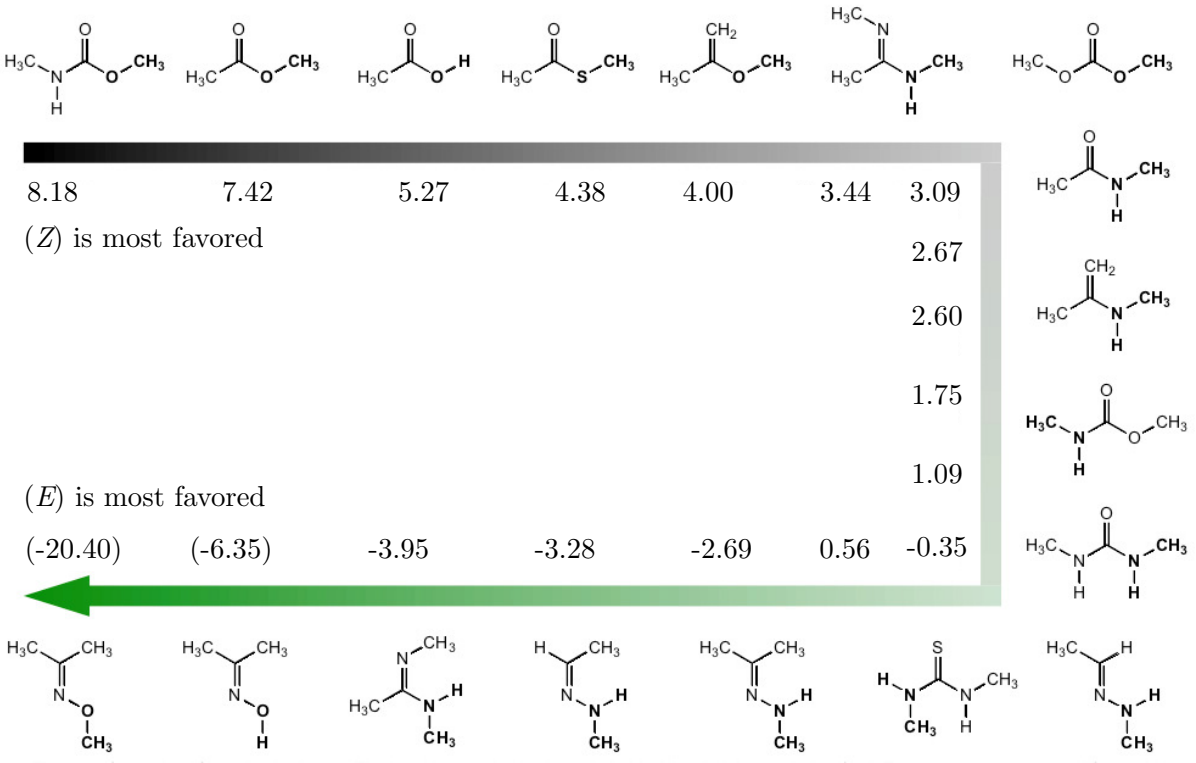


Figure 2.5: Summary of the G3B3 E/Z free-energy differences (kcal/mol). The preferred conformation is shown, and the fragment that is rotated is highlighted in bold.

2.2.3 GB/SA results.

The gas-phase results for the E/Z preferences can be shifted in different molecular environments, both relatively homogeneous as for a pure solvent and inhomogeneous as in a protein binding site. To gain some sense of magnitude for the former case, free energies of hydration were calculated for all conformers using the OPLS/CM1A force field and GB/SA continuum solvent model.[34,90] The gas-phase G3B3 results, the GB/SA shifts $\Delta\Delta G_{\mathrm{hyd}}$, and the net ΔG_{aq} for the conformational equilibria in aqueous solution at 298 K are summarized in Table 2.4. A negative G_{hyd} indicates that the E conformer is predicted to be better hydrated. Based on calculations for 399 neutral organic molecules, the average absolute error for free energies of hydration from the GB/SA calculations is expected to be 1.0 kcal/mol.[35] The errors for the differential hydration of conformers should be smaller, and results for several standard cases were shown to be in good accord

with experimental data.[35] However, E/Z conformers may be particularly challenging owing to the accompanying changes in solute–water hydrogen bonding as compared to simpler cases such as the *gauche* $\rightleftharpoons$ *anti* equilibria for 1,2-dihaloethanes.[35]

Table 2.4: Computed $E - Z$ free-energy (kcal/mol) and dipole (D) differences in the gas phase and in aqueous solution at 298 K.

Pair	ΔG_{gas}	$\Delta\mu$	$\Delta\Delta G_{\text{hyd}}$	ΔG_{aq}
7	8.18	3.05	−0.52	7.66
3	7.42	3.10	−0.60	6.82
1	5.27	2.92	−2.39	2.88
4	4.38	3.19	−0.12	4.26
11	4.01	1.27	−1.82	2.18
13	3.44	−0.04	1.75	5.19
9	3.09	3.60	0.26	3.35
2	2.67	0.32	1.43	4.10
10	2.60	−0.13	0.29	2.89
8	1.75	0.34	0.37	2.12
5	1.09	0.50	0.00	1.09
14	−0.35	−0.01	−1.02	−1.37
6	−0.55	0.80	1.28	0.73
16	−2.69	0.54	−1.06	−3.75
15	−3.28	0.37	−0.36	−3.64
12	−3.95	0.31	1.79	−2.16
(17)†	−6.35	−2.94	−0.11	−6.46
(18)†	−20.40	−2.66	−1.40	−21.80

†The planar Z form **b** is a transition state.

The computed $\Delta\Delta G_{\text{hyd}}$ values in Table 2.4 fall in a relatively narrow range, ± 2 kcal/mol, so the shifts were generally not found to be enough to qualitatively change the direction of the E/Z equilibria. The possible exception was for thiourea (pair **6**), for which ΔG was found to be 0 ± 1 kcal/mol in all media.[74] The expectation from classical electrostatics is that, in the absence of steric effects, the conformer with the larger dipole moment should have a more negative free energy of hydration.

Thus, for most cases in Tables 2.1 and 2.2, the E conformer was found to be better hydrated than the Z conformer. In this regard, the results in Table 2.4 were mixed. For acetic acid (pair **1**) the E conformer was found to have a 2.92 D larger dipole moment

than the Z form and it was found to be better hydrated by 2.39 kcal/mol. This value is significantly smaller in magnitude than estimates of $\Delta\Delta G_{\mathrm{hyd}}$ from a QM/MM study in TIP4P water (−4.8 kcal/mol)[99] and from QM/RISM calculations (−5.2 kcal/mol).[100] If these values are combined with the G3B3 gas-phase result, the prediction is that (E)- and (Z)-acetic acid are nearly equally populated in water at 298 K or, equivalently, that the Brønsted basicities of the *syn* and *anti* lone pairs for acetate ion in water would be similar.[101,102] It should be noted that in dilute aqueous solution at neutral pH, less than 1% of acetic acid is not ionized.

Furthermore, the ester in pair **3** and the carbamate in pair **7** also were found to have changes of ca. 3 D in dipole moment, but the E conformer was predicted to be better hydrated by only ca. 0.6 kcal/mol. Previous results for pair **3** from free energy perturbation calculations in TIP4P explicit water predicted preferential hydration of the E conformer by 3.0 kcal/mol.[69] Most surprisingly, although the E,Z conformer of the carbonate (pair **9**) has a 3.60 D larger dipole moment than the Z,Z conformer, the Z,Z conformer was predicted to be better hydrated by 0.26 kcal/mol. The results for N-methylacetamide (pair **2**) also appear to be off the mark. There is general consensus that the E/Z equilibrium for N-methylacetamide is affected little by hydration,[64,83] while the GB/SA results were found to favor hydration of the Z conformer by 1.43 kcal/mol. The differential hydration arises predominantly from differences in the GB term; the SA term varies by less than 0.1 kcal/mol for these E/Z equilibria.

The noted discrepancies do not reflect obvious problems with the 1.07*CM1A charges that are used in the GB/SA calculations. The computed dipole moments with these charges mimic the G3B3 results well. For instance, the 1.07*CM1A dipole moments for (E,Z)- and (Z,Z)-dimethylcarbonate were found to be 3.70 and 0.44 D, which are close to the G3B3 values of 3.97 vs 0.37 D. And, for (E)- and (Z)-N-methylacetamide, the 1.07*CM1A dipole moments were 4.00 and 3.44 D, while the G3B3 results were 4.54 and 4.22 D. Further examination of solvent effects on the E/Z equilibria is warranted

using free-energy methods in simulations with explicit solvent. In view of the expected sensitivity of the results to details of solute–solvent hydrogen bonding, it is unclear if continuum models can accurately gauge solvent effects in such cases.

2.3 Summary and conclusions.

Changes in energy, enthalpy, free energy, and dipole moment were evaluated at the G3B3 level for 18 pairs of conformers exhibiting prototypical E/Z conformational equilibria for rotation about single bonds. The results are important for consideration in molecular design and in the evaluation of structures that arise from protein–ligand docking studies as well as from crystallography. For the systems studied, which included representatives of carboxylic acids, carboxylic esters, thioesters, secondary amides, ureas, carbamates, carbonates, enol ethers, enamines, and amidines, the preferred conformer is normally Z. Preference for the E conformer mostly arises from steric effects in hydrazones, amidines, and oximes that destabilize the Z conformer, especially via *syn*-1,5-CC interactions. A particularly interesting case is 1,3-dimethylthiourea, which is found to slightly favor the E,Z conformer over the Z,Z alternative in the gas phase. Free energies of hydration were also estimated for the conformers from GB/SA calculations. Accurate computation of the effects of hydration on E/Z equilibria is expected to be particularly challenging in view of the substantial, accompanying changes in solute–water hydrogen bonding. Though the differential effects from the GB/SA calculations were generally found to be insufficient to overcome the gas-phase preferences, the computed effects in several cases seem too small. Further investigation is warranted with free-energy methods in molecular dynamics or Monte Carlo simulations using explicit hydration to obtain more accurate results and to provide a basis for testing and improvement of continuum solvation methods.

Chapter 3

OPLS torsion profiles for derivatives of drug-like heterocycles.

3.1 Introduction

Well-defined molecular mechanics are central to applications in computer-aided drug design, and the insights they afford extend into other areas of chemistry including physical chemistry, organic synthesis, and medicinal chemistry.[61] Of particular interest to those in the field of molecular design is the ensemble of conformations defined by one or more dihedral torsions within a given molecule. Often it is the case that these torsions give rise to unique molecular shapes, which can affect a potential drug's ability to bind (or not bind) to its intended target. In efforts to develop novel and potent non-nucleoside reverse transcriptase inhibitors (NNRTIs) targeting HIV-1 reverse transcriptase, we have previously encountered a number of highly active drug leads containing benzimidazole,[103] 2-furanyl,[104] 2-pyrimidinyl,[105] oxazole, and oxadiazol[60] cores and motifs, while others have reported similar findings for leads containing 2-thiophenyl motifs[106] and pyrrole cores.[107] The need for well-parameterized force fields as they relate to conformational energetics of small organic molecules has previously been highlighted,[108] especially in biomolecular systems where quantum mechanical calculations are unfeasible. Others have reported on the importance of dihedral torsions in this regard.[61] To this end, we sought new torsion parameters in the OPLS force field[22] for 65 prototypical derivatives of benzene, pyridine, furan, thiophene, and pyrrole, containing methyl, ethyl, isopropyl, cyclopropyl, *t*-butyl, vinyl, hydroxy, methoxy, thio, methylthio, amino, *N*-methylamino,

and *N*,*N*-dimethylamino substituents at the 1- (benzene) or 2- (pyridine, furan, thiophene, and pyrrole) position of the ring, Figure 3.1. The new parameters were developed by fitting torsion profiles to quantum mechanical data and represent fundamental molecular mechanics that will support current and future endeavors in state-of-the-art drug design methodology.

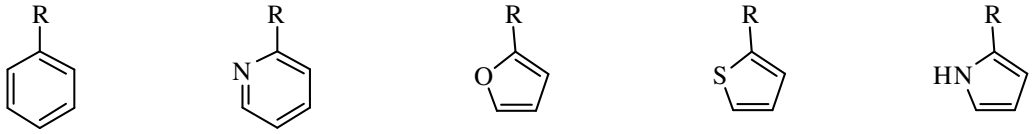


Figure 3.1: The five heterocyclic cores for which torsion parameters were developed: benzene, 2-pyridine, 2-furan, 2-thiophene, and 2-pyrrole, where R = CH_3, Et, iPr, cPr, tBu, vinyl, OH, OCH_3, SH, SCH_3, NH_2, $NHCH_3$, and $N(CH_3)_2$.

Monte Carlo simulations have been used for years to understand a variety of chemical and biochemical phenomena[109] ranging from the properties of pure liquids to the intricate details of protein–ligand interactions. The former has in fact long been the basis for many OPLS-AA atom types, utilizing pure liquid simulations to seed and refine atomic charges and Lennard–Jones parameters.[28,110] For the purposes of this work, semi-empirical CM1A atomic charges scaled by 1.14 were chosen over OPLS-AA atomic charges due to their flexibility and ease of application, which often make CM1A the atomic charge model of choice for large biomolecular systems and therefore the more relevant charge model for our purposes here.

Among a myriad of other uses, the ensemble of states generated during a Monte Carlo simulation can also be exploited to give angle and dihedral distributions throughout a population, denoted $S(\phi)$ for dihedrals. Furthermore, since Monte Carlo simulations can in theory be performed in any solvent as well as in the gas phase, insight can be gained into how a dihedral angle distribution is affected by solvation (assuming that the solvent is properly parameterized as well), most notably with water. Thus, dihedral angle distributions for the key torsions in the gas phase and in water were used to assess the performance of the new parameters, as well as to help elucidate the behavior of the

torsions in a more biochemically relevant environment. Post-parameterization torsion profiles were also generated using the generalized Born / surface area (GB/SA) implicit solvation model of Still and co-workers[27,28] with the goal of reinforcing results from the Monte Carlo simulations.

3.1.1 Computational details.

3.1.1.1 Ab initio calculations.

All ab initio quantum mechanical calculations were carried out at the MP2/6-311G(d) level of theory using the `GAUSSIAN03` program.[91] Quantum mechanical (QM) torsion profiles were determined via a relaxed potential-energy surface scan about the key dihedral. Scans were performed for 18 increments of $\phi = 10^\circ$ with redundant coordinates and the resulting curves were extrapolated to 360° based on molecular symmetry. For the amino derivatives, additional improper dihedrals were used to prevent inversion of the amino nitrogens, ensuring smooth torsions.

3.1.1.2 Force field calculations.

All force field calculations were performed using the *BOSS* program,[33] which calculates the total potential energy of a system as a function of harmonic bond stretching and angle bending terms, a Fourier series for each dihedral torsion, and Coulomb and Lennard–Jones terms for nonbonded (1,¿4) interactions, eqs. 3.1–3.5. Full molecular mechanical (MM) torsion profiles were computed in the gas phase and with the generalized Born / surface area (GB/SA) solvation model as implemented in *BOSS*. Of specific interest were the values of the torsion parameters V_1, V_2, V_3, and V_4 in the Fourier series for each dihedral type. Structures were optimized to an energy minimum using the OPLS/CM1A force field with CM1A charges scaled by 1.14, and the gas phase torsion parameters were fit to the MP2 quantum mechanical data.

$$E_{\text{total}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}} + E_{\text{non-bonded}} \tag{3.1}$$

$$E_{\text{bonds}} = \sum_{\text{bonds}} K_r(r - r_0)^2 \tag{3.2}$$

$$E_{\text{angles}} = \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 \tag{3.3}$$

$$E_{\text{torsions}} = \sum_{\text{torsions}} \sum_{N=1}^{4} \left\{ V_N \left(\frac{1 + \cos(N\phi)}{2} \right) \right\} \tag{3.4}$$

$$E_{\text{non-bonded}} = \sum_i \sum_j \left\{ \frac{q_i q_j e^2}{r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{3.5}$$

3.1.1.3 Monte Carlo simulations.

Monte Carlo simulations with Metropolis sampling were performed for each derivative in the gas phase and in TIP4P liquid water at 25 °C and 1 atm pressure. All molecules were fully flexible; no geometric constraints were imposed. For the gas phase simulations, statistics were averaged over 2.0×10^6 moves after an initial 1.0×10^6 moves of equilibration. For the TIP4P simulations, a periodic box containing 512 water molecules was equilibrated with the solute for 2.0×10^6 moves in the NPT ensemble after an initial 200K moves of NVT equilibration. Statistics were then averaged for an additional 1.0×10^7 moves. During sampling, key dihedrals were "flipped" at random intervals by 180° to attempt a more complete exploration of conformational space for molecules whose conformers were energetically separated by substantial potential-energy barriers. Key dihedral angle distributions in both gas and condensed-phase environments were evaluated from the resulting data.

3.2 Experimental design and results.

3.2.1 Definition and development of torsion parameters.

The development of torsion parameters for new dihedral types in the OPLS force field was an iterative process involving simultaneous optimization of V_1, V_2, V_3, and V_4 for key dihedrals in each core/fragment combination in Figure 3.1. For molecules in the benzene series, the key dihedrals were defined from the 2-carbon in the ring to each 1,4-related atom type in the R group. For instance, in phenol, one key dihedral was defined: HO–OH–A–CA. For *N*-methylaniline, two key dihedrals were defined: H–N–CA–CA and CT–N–CA–CA. A graphical representation of these two examples is shown in Figure 3.2.

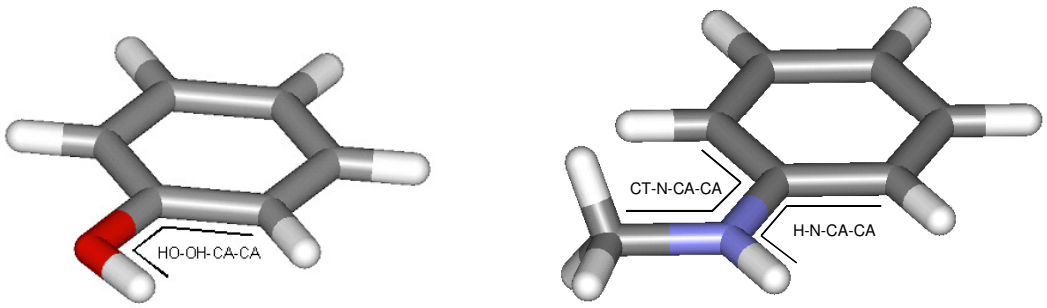


Figure 3.2: Graphical representation of the key dihedrals defined phenol (left) and *N*-methylaniline (right) from the benzene series.

For all other molecules (i.e., those in the pyridine, furan, thiophene, and pyrrole series), the key dihedrals were defined from the heteroatom in the ring to each 1,4-related atom type in the R group, and from the 3-carbon in the ring to each 1,4-related atom type in the R group. For instance, in 2-hydroxyfuran, two key dihedrals were defined: HO–OH–CW–OS and HO–OH–CW–CS, and in 2-cyclopropylfuran, four key dihedrals were defined: HC–CY–CW–OS, HC–CY–CW–CS, CY–CY–CW–OS, and CY–CW–CW–CS. A graphical representation of these two examples is shown in Figure 3.3.

For each dihedral type, values for V_1, V_2, V_3, and V_4 were optimized to give a torsion profile most closely resembling that which was obtained through MP2 quantum mechanical

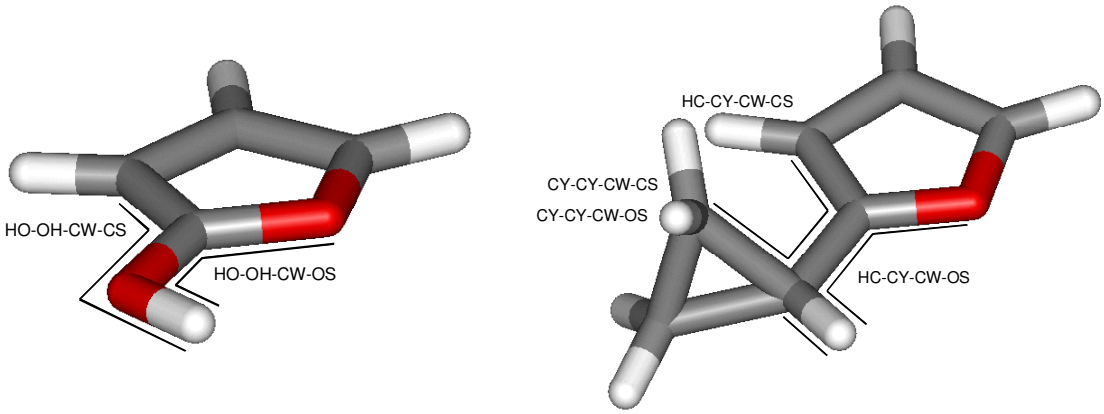


Figure 3.3: Graphical representation of the key dihedrals defined in 2-hydroxyfuran (left) and 2-cyclopropylfuran (right).

calculations (taken here to be the reference). A complete list of all the dihedral types and their optimized parameters is given in the Appendix. Significant improvements were achieved for most molecules in this set. To give a quantitative sense of this improvement, the mean unsigned errors (MUEs) between all scan points and their reference counterparts were averaged for all molecules in each series, both before parameterization and after parameterization, Table 3.1.

Table 3.1: Quantitative comparison of the mean unsigned error (MUE) between all data points and their reference counterparts both before and after parameterization. All values are in kcal/mol.

Series	MUE Before	MUE After	Reduction of Error
Benzenes	0.53	0.18	66.9%
Pyridines	1.40	0.34	75.4%
Furans	2.24	0.12	94.6%
Thiophenes	3.25	0.14	95.8%
Pyrroles	2.87	0.18	93.6%

As can be seen in Table 3.1, the benzene series showed the least improvement, owing to a more completely defined parameter set before the start of this project. Nevertheless, a nearly 67% reduction of error was achieved between the MM and QM torsions in the benzene series, with the average error at each scan point being reduced from 0.53 kcal/mol to 0.18 kcal/mole. A similar improvement was seen within the pyridine series. Of note

were the furan, thiophene, and pyrrole series, which benefited the most from the inclusion of new dihedral types and finely tuned parameters. For the furan series, the average error at each scan point was reduced from 2.24 kcal/mol to 0.12 kcal/mol, a 94.6% reduction in error. The average error at each scan point was reduced from 3.25 kcal/mol to 0.14 kcal/mol (95.8%) within the thiophene series and from 2.87 kcal/mol to 0.18 kcal/mol (93.6%) within the pyrrole series. Examples from each series follow. In each case, the MM curve was computed with the OPLS/CM1A force field and the QM curve was computed with at the MP2/6-311G(d) level of theory.

Figure 3.4 shows the MM and QM torsion profiles for phenol, a constituent of the benzene series, both before and after parameterization. Before parameterization (Figure 3.4, top) the MM torsion profile exhibited a smooth, well-behaved curve similar in shape to the QM torsion profile, but with underestimated energy barriers at 90/270°. Modest reparameterization of the key dihedral (HO–OH–CA–CA) yielded much more accurate energy barriers (Figure 3.4, bottom). Figure 3.5 shows the MM and QM torsion profiles for 2-vinylpyridine, a constituent of the pyridine series, both before and after parameterization. Before parameterization (Figure 3.5, top) the MM torsion profile exhibited a smooth, well-behaved curve similar in shape to the QM torsion profile, but with severely overestimated energy barriers at 90/270°. Parameterization of the key dihedrals (CM–C=–CA–NC, CM–C=–CA–CA, HC–C=–CA–NC, and HC–C=–CA–CA) yielded a nearly perfectly matched curve, with appropriate energy barriers, local and global minima (Figure 3.5, bottom). Note that two of the key dihedrals for 2-vinylpyridine (CM–C=–CA–NC and HC–C=–CA–CA) are shared with, and had already been optimized for, styrene (vinylbenzene) from the benzene series. Figure 3.6 shows the MM and QM torsion profiles for 2-(methylthio)furan, a constituent of the furan series, both before and after parameterization. Before parameterization (Figure 3.6, top) the MM torsion profile exhibited a mostly smooth, well-behaved curve, however its shape was inverted compared to the QM torsion profile and it yielded severely overestimated energy barriers.

Parameterization of the key dihedrals (CT–S–CW–CS and CT–S–CW–OS) yielded a nearly perfectly matched curve, only underestimating the energy barrier at 0/360° by approximately 0.25 kcal/mol. Most notable in this example is the formation of the correct saddle point around 180° (Figure 3.6, bottom).

Figure 3.7 shows the MM and QM torsion profiles for *N*-methylthiophen-2-amine, a constituent of the thiophene series, both before and after parameterization. Before parameterization (Figure 3.7, top) the MM torsion profile exhibited a curve not resembling its QM counterpart in most ways. Indeed the amino derivatives proved to be the most problematic to model: curves were generally more erratic and responded less systematically to changes of the torsion parameters. Careful parameterization of the key dihedrals (CT–N–CW–CS, CT–N–CW–S, H–N–CW–CS, and H–N–CW–S) afforded a curve that more closely resembles the QM torsion profile: saddle points at 0/360° and 180° are represented, although the relative energies differ by 0.1–0.5 kcal/mol, Figure 3.7, bottom. Figure 3.8 shows the MM and QM torsion profiles for 2-cyclopropylpyrrole, a constituent of the pyrrole series, both before and after parameterization. Before parameterization (Figure 3.8, top) the MM torsion profile exhibited a mostly smooth, well-behaved curve; however, as was the case for 2-(methylthio)furan, its shape was inverted compared to the QM torsion profile and it yielded severely overestimated energy barriers. Parameterization of the key dihedrals (HC–CY–CW–N2, HC–CY–CW–CS, CY–CY–CW–N2, CY–CY–CW–CS) yielded a nearly perfectly matched curve, with the appropriate saddle points, energy barriers, local and global minima. Most notable in this example is the formation of the correct saddle point around 180°, Figure 3.8, bottom.

3.2.2 Monte Carlo simulations and torsion profiles with GB/SA solvation.

Pleased with the effects of the parameterization on reproducing torsion profiles, it was of particular interest to study the behavior of the newly defined dihedral types in standard Monte Carlo simulations. Such simulations allow for insight into specific dihedral angle

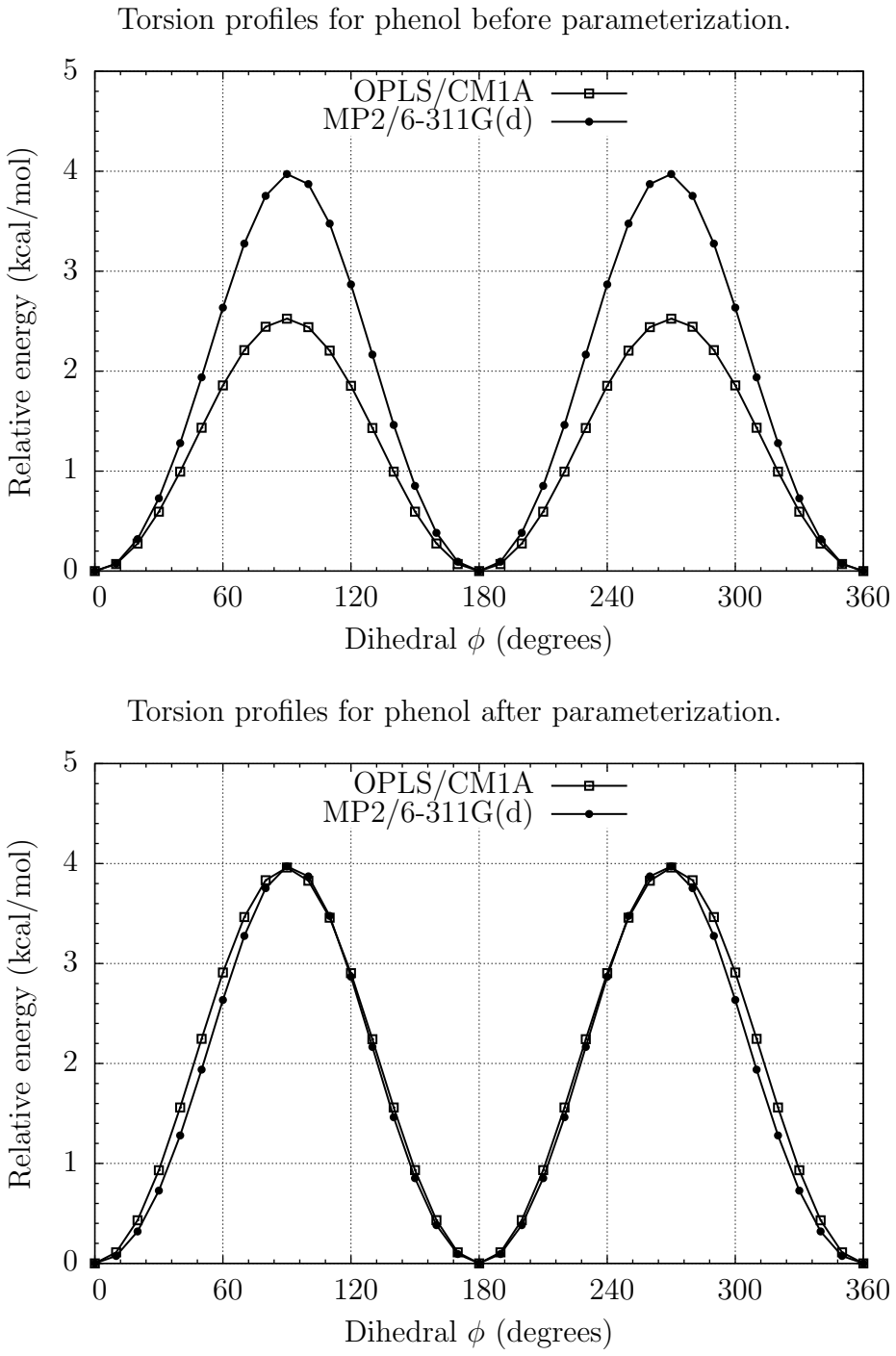
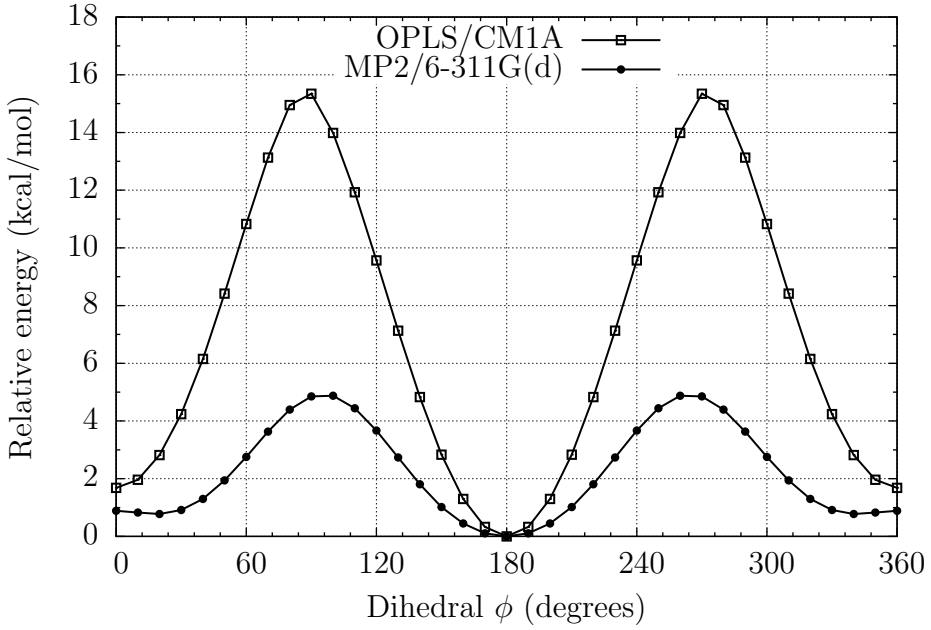


Figure 3.4: The MM (OPLS/CM1A) and QM (MP2/6-311G(d)) torsion profiles for phenol, a constituent of the benzene series, both before (top) and after (bottom) parameterization. The 0° dihedral was taken to be the H–O–C1–C2 eclipsed conformer.

Torsion profiles for 2-vinylpyridine before parameterization.



Torsion profiles for 2-vinylpyridine after parameterization.

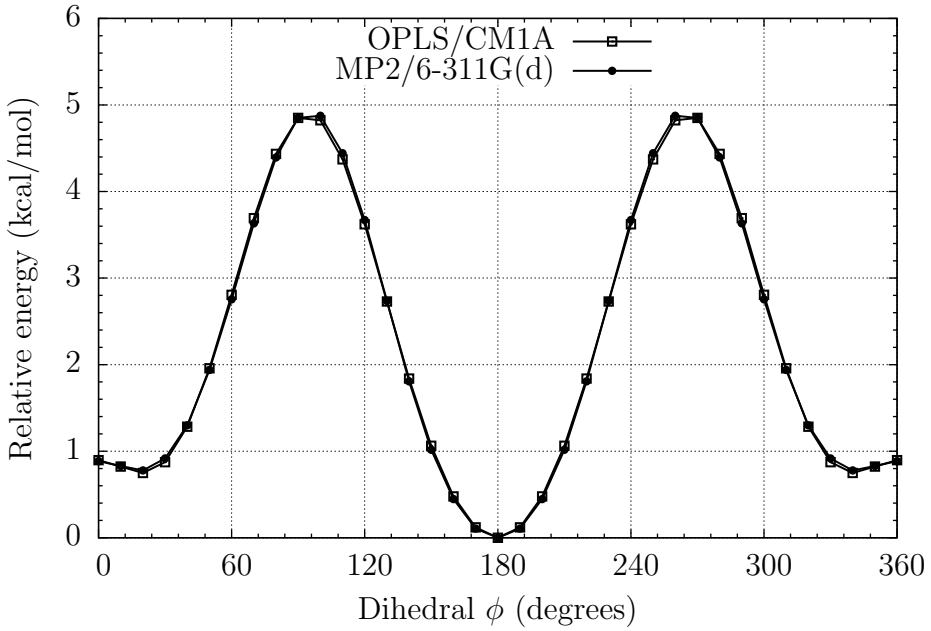
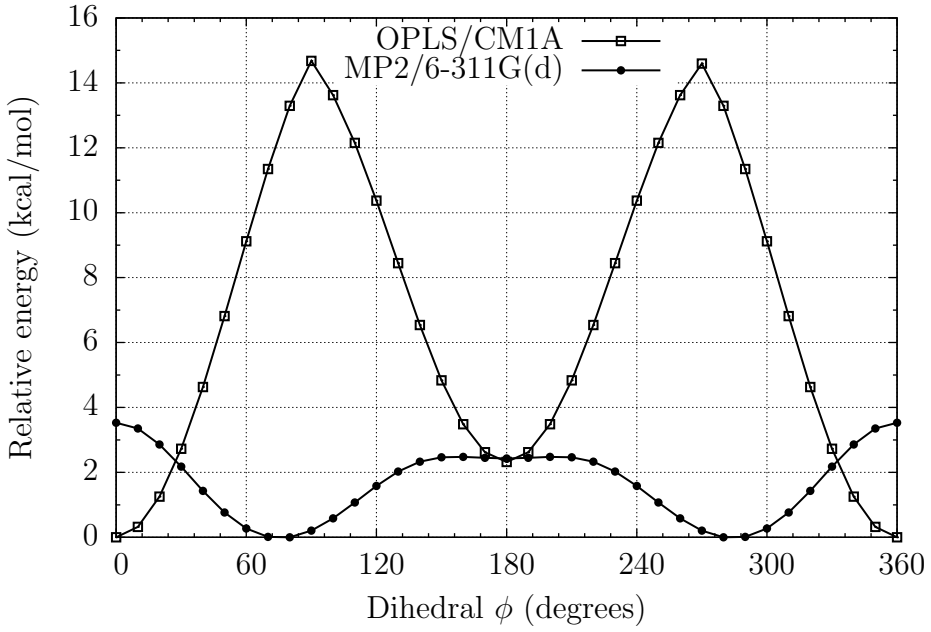


Figure 3.5: The MM (OPLS/CM1A) and QM (MP2/6-311G(d)) torsion profiles for 2-vinylpyridine, a constituent of the pyridine series, both before (top) and after (bottom) parameterization. The 0° dihedral was taken to be the CM–C=–C2–N eclipsed conformer.

Torsion profiles for 2-(methylthio)furan before parameterization.



Torsion profiles for 2-(methylthio)furan after parameterization.

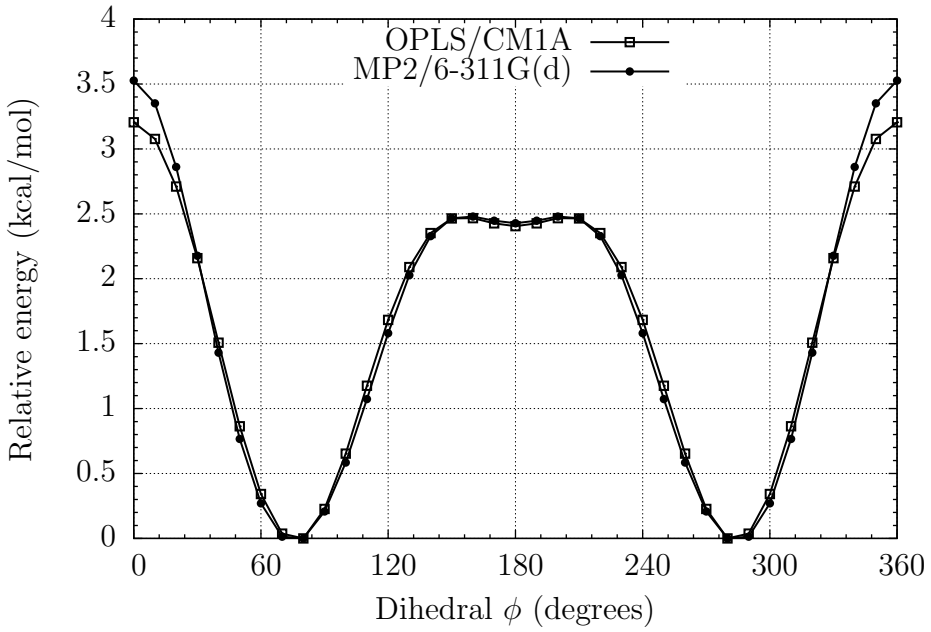
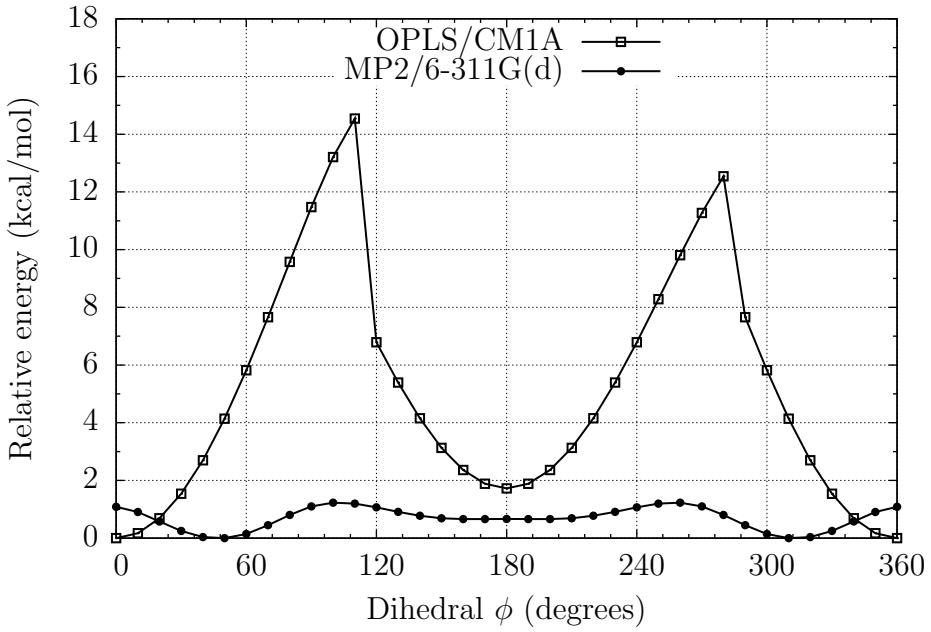


Figure 3.6: The MM (OPLS/CM1A) and QM (MP2/6-311G(d)) torsion profiles for 2-(methylthio)furan, a constituent of the furan series, both before (top) and after (bottom) parameterization. The 0° dihedral was taken to be the CT–S–C2–O eclipsed conformer.

Torsion profiles for N-methylthiophen-2-amine before parameterization.



Torsion profiles for N-methylthiophen-2-amine after parameterization.

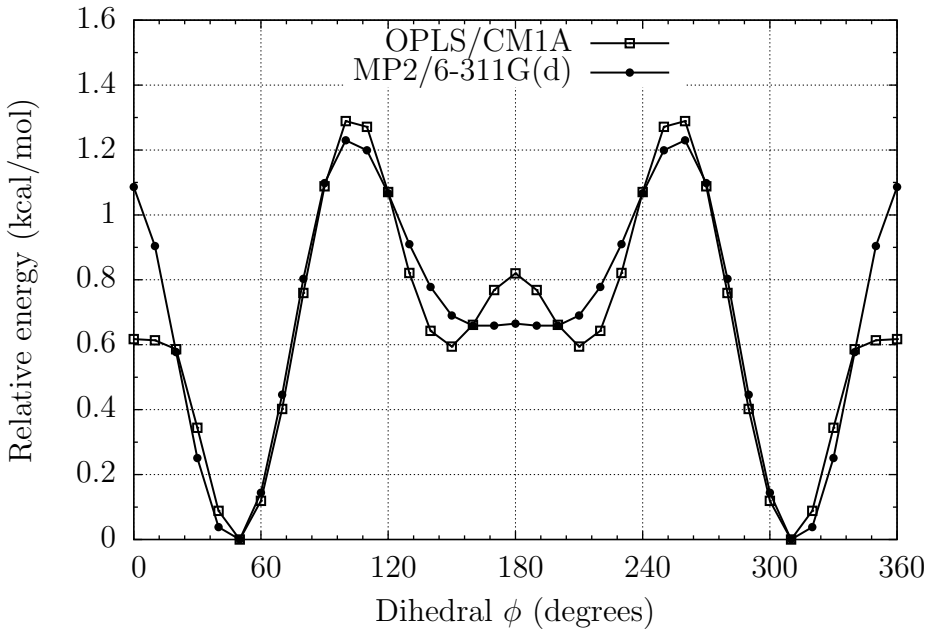
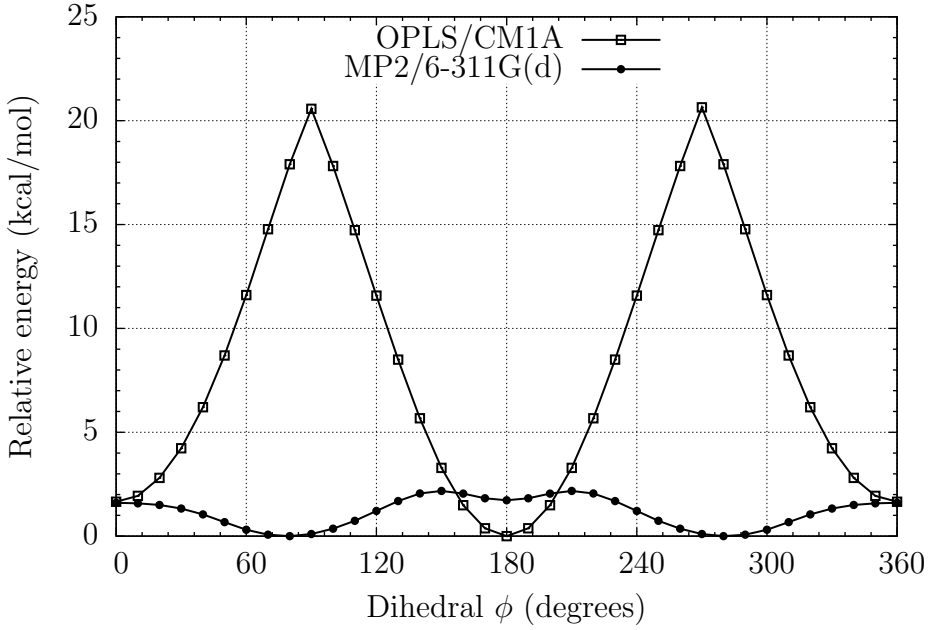


Figure 3.7: The MM (OPLS/CM1A) and QM (MP2/6-311G(d)) torsion profiles for N-methylthiophen-2-amine, a constituent of the thiophene series, both before (top) and after (bottom) parameterization. The 0° dihedral was taken to be the H–N–C2–S eclipsed conformer.

Torsion profiles for 2-cyclopropylpyrrole before parameterization.



Torsion profiles for 2-cyclopropylpyrrole after parameterization.

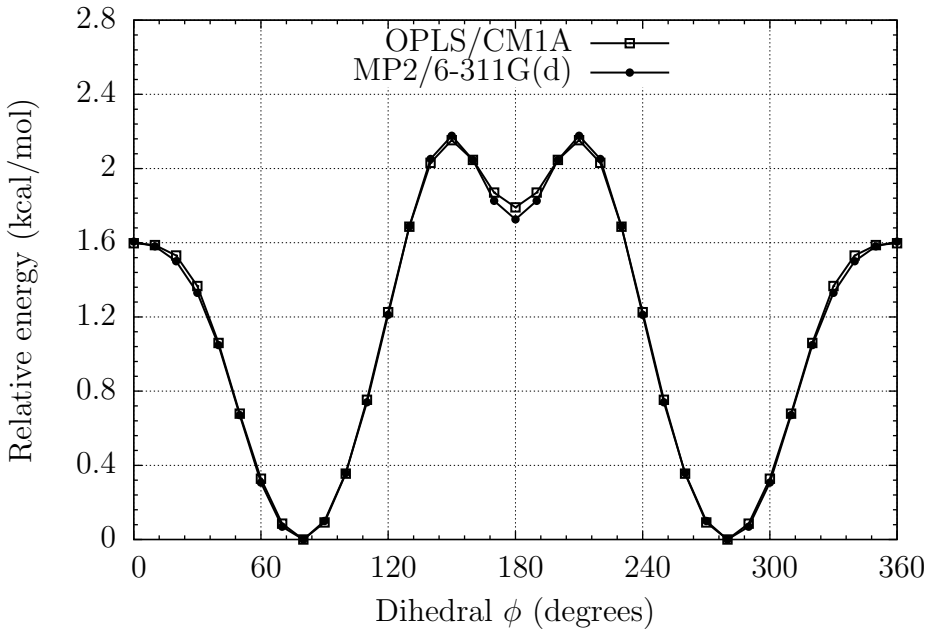


Figure 3.8: The MM (OPLS/CM1A) and QM (MP2/6-311G(d)) torsion profiles for 2-cyclopropylpyrrole, a constituent of the pyrrole series, both before (top) and after (bottom) parameterization. The 0° dihedral was taken to be the HC–CY–CW–N eclipsed conformer.

distributions within a population; thus, given a well-defined torsion profile, the dihedral angle distribution observed in a Monte Carlo ensemble should mirror the relative energies in the torsion profiles. Keeping in mind that the motivation for this work was to improve efforts in drug design, it was also of interest to understand if any shift in dihedral angle distribution could be observed when the solute is moved from an isolated environment (gas phase) to an aqueous one, as discussed in the previous chapter and not unlike what one would encounter in a biomolecular system. Furthermore, any shift in the observed dihedral angle distribution in the aqueous Monte Carlo ensemble should be reproducible in a torsion scan using GB/SA implicit solvation.

It was encouraging to see that the torsion profiles were reflected in the gas-phase dihedral angle distribution $S(\phi)$ for all 65 molecules studied. As an example of this, both the dihedral angle distribution $S(\phi)$ and the torsion profile for thiophene-2-thiol are overlaid in Figure 3.9. For thiophene-2-thiol, the key dihedral was HS–SH–CW–S, and the 0° dihedral was taken to be the HS–SH–CW–S eclipsed conformer. Several things are particularly noteworthy in Figure 3.9. First, an essentially perfect agreement between the QM and MM torsion profiles is seen. From these torsion profiles one can learn that a moderate barrier to rotation of about 3.5 kcal/mol, with minima at 80° and 280° and maxima at 0/360° and 180° was predicted. Second, the gas-phase dihedral angle distribution $S(\phi)$ was indeed found to mirror the gas-phase torsion profile. The bulk of the population was found to be clustered around angles that give energy minima, and almost zero population was found at 0/360° and 180°, which coincide with rotational energy barriers. Third, it is clear that $S(\phi)$ was unchanged between gas phase and aqueous environments. This can be reasoned by consideration of both the rotational energy barrier and the change in dipole between the major conformers. For thiophene-2-thiol, change in dipole was found to be only modest compared to the rotational energy barrier, ranging from 1.08 D at 0° to 1.75 D at 80° to 2.42 D at 180°; thus, one would not expect a significant change in dihedral angle distribution between gas and aqueous environments.

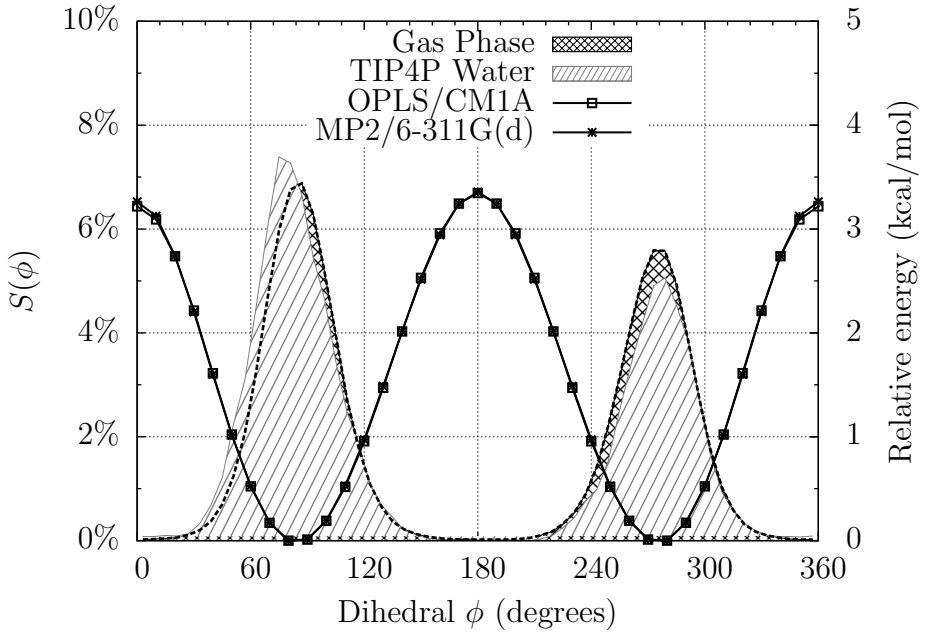


Figure 3.9: Dihedral angle distribution $S(\phi)$ and torsion profile overlaid for thiophene-2-thiol.

Only eight of the 65 molecules that were studied exhibited a significant shift in dihedral angle distribution between gas and aqueous environments. The five most prominent of these unique cases are summarized in Table 3.2. In summary of other cases such as the methyl, ethyl, isopropyl and *tert*-butyl derivatives, no significant shift was observed, owing to small (ca. 0.6–2 kcal/mol) rotational energy barriers where all conformations could be easily achieved regardless of environment, negligible changes in dipole between the different conformers, and/or insignificant changes in local minima, diminishing any preference for one conformer over the other in a given environment. For cyclopropyl and vinyl derivatives with the exception of 2-vinylpyrrole, no population shift was observed, owing to larger (ca. 3–6 kcal/mol) rotational energy barriers and negligible changes in dipole between the different conformers. In the case of 2-vinylpyrrole, an increased rotational energy barrier and lower-energy local minimum in water afforded a shift away from the less polar global minimum at 0° (CM–C=–CW–N2) towards the more polar 180° local minimum. For the amino derivatives with the exception of *N*-methylthiophen-2-amine

and *N*-methyl-1*H*-pyrrol-2-amine, no population shifts were observed either, owing to large rotational energy barriers and the absence of local minima or saddle points. As was the case for 2-vinylpyrrole, in *N*-methylthiophen-2-amine an increased rotational energy barrier and lower-energy local minimum in water afforded a shift away from the less polar global minima at 50/310° (H–N–CW–S) towards the more polar 180° local minimum. For *N*-methyl-1*H*-pyrrol-2-amine, a more modest shift was observed from the less polar global minima at 60/300° (H–N–CW–N2) to the more polar local minima at 0°, owing to less pronounced torsion profile.

Table 3.2: Summary of the most unique cases where significant dihedral angle distribution shift was observed between gas and aqueous environments. GM = global minimum dihedral angle, LM = local minimum dihedral angle. Angles are in degrees, dipoles are in Debye and energies are in kcal/mol.

R	GM	μ_{GM}	LM	μ_{GM}	$\Delta\mu$	$E_{\mathrm{gas}}^{\mathrm{Barrier}}$	$E_{\mathrm{GB/SA}}^{\mathrm{Barrier}}$
Pyridines							
OH	0	1.8	180	4.59	2.79	10.22	6.7
SH	0	1.86	180	3.63	1.77	3.76	Inverted
SCH_3	0	1.11	180	4.00	2.89	3.8	2.5
Furans							
OH	0	1.40	180	2.8	1.4	1.94	Inverted
OCH_3	180	2.75	60/300	1.8	−0.95	2.08	2.7

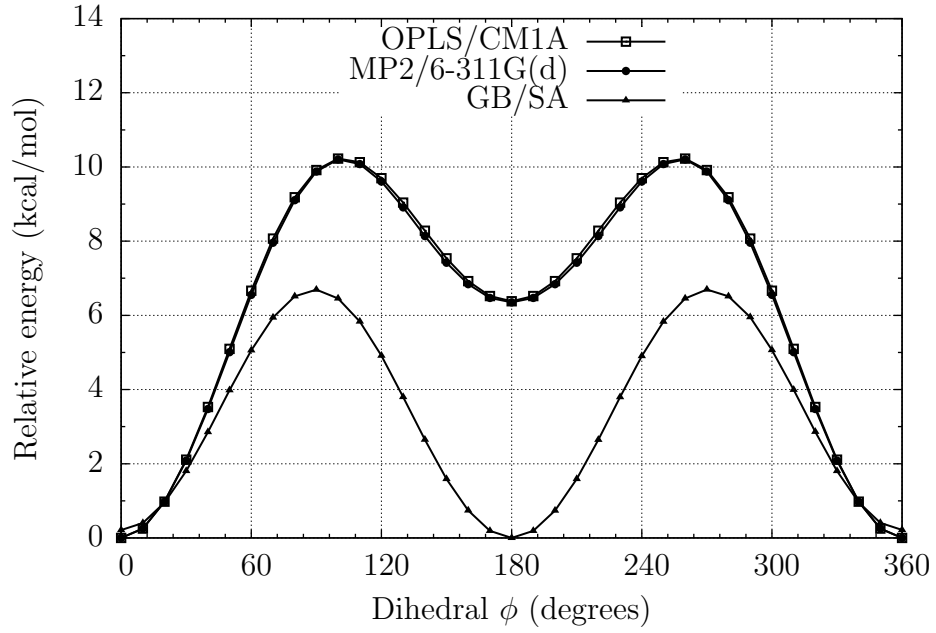
In analyzing the prominent cases summarized in Table 3.2, we first turn to 2-hydroxypyridine, which exhibited one of the highest gas-phase rotational energy barriers of all the molecules studied, at 10.22 kcal/mol. The torsion profile and dihedral angle distributions are shown in Figure 3.10. The global minimum for 2-hydroxypyridine was found at a HO–OH–C–N dihedral angle of 0° (eclipsed). A local minimum was found at 180°, with a relative energy of approximately 6.4 kcal/mol and an intervening energy barrier of 10.22 kcal/mol between global and local minima. The change in dipole between global and local minima was moderate at 2.79 D. As can be seen in the dihedral angle distributions (Figure 3.10, bottom), a drastic shift in population was found to occur between the gas phase and aqueous environments. The gas phase population mirrored

the gas phase torsion profile, with the population clustered tightly around the less polar (1.8 D) global minimum of 0/360°. In water, the population was found to shift exclusively to the more polar (4.59 D) local minimum of 180°. Indeed, this shift is reflected in the GB/SA torsion profile, which shows a marked deviation from that of the gas phase, lowering the energy barrier to around 6.7 kcal/mol and reducing the ΔE between the global and local minima to 0 kcal/mol.

Turning to pyridine-2-thiol we find a similar situation, though despite exhibiting a much smaller rotational energy barrier than 2-hydroxypyridine, the effects were found to be equally if not more dramatic. The torsion profile and dihedral angle distributions are shown in Figure 3.11. The global minimum for pyridine-2-thiol was found at a HS–SH–C–N dihedral angle of 0° (eclipsed). A local minimum was found at 180°, with a relative energy of approximately 1.6 kcal/mol and an intervening rotational energy barrier of 3.76 kcal/mol between global and local minima. The change in dipole between global and local minima was modest at 1.77 D. As can be seen in the dihedral angle distributions (Figure 3.11, bottom), again, a drastic shift in population was found to occur between the gas phase and aqueous environments. The gas phase population mirrored the gas phase torsion profile, with the population clustered tightly around the less polar (1.86 D) global minimum of 0/360°. In water, the population shifted exclusively to the more polar (3.63 D) local minimum of 180°. This shift is once again reflected in the GB/SA torsion profile. In fact, unlike for 2-hydroxypyridine where the global and local minima were equalized, for pyridine-2-thiol the global and local minima were found to have flipped, or inverted, with the more polar 180° conformer becoming the new global minimum and the less polar 0° conformer becoming the new local minimum, with a ΔE approximately equal to that of the gas phase and a slightly larger intervening rotational energy barrier of around 3.9 kcal/mol.

The gas phase and GB/SA torsion profiles for 2-(methylthio)pyridine were found to resemble those of 2-hydroxypyridine but with a smaller energy barrier and ΔE, Figure 3.12,

Torsion profile for 2-hydroxypyridine.



Dihedral angle distributions for 2-hydroxypyridine.

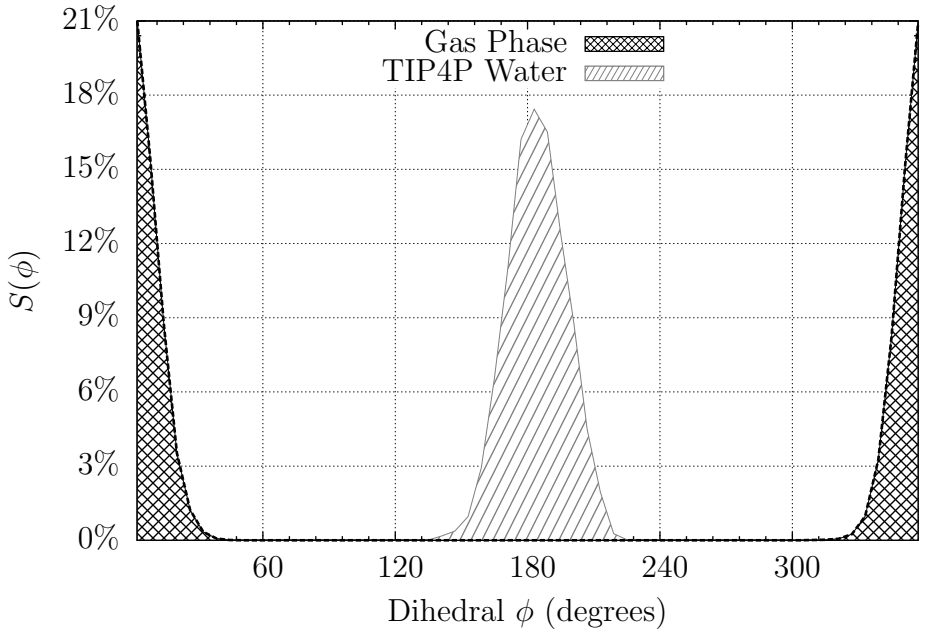
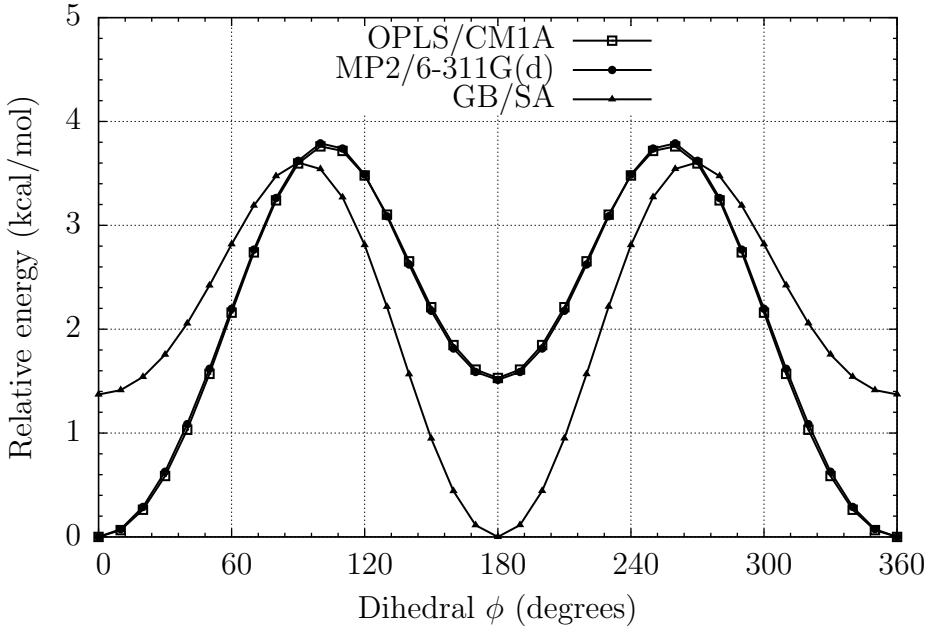


Figure 3.10: Torsion profile and dihedral angle distributions for 2-hydroxypyridine.

Torsion profile for pyridine-2-thiol.



Dihedral angle distributions for pyridine-2-thiol.

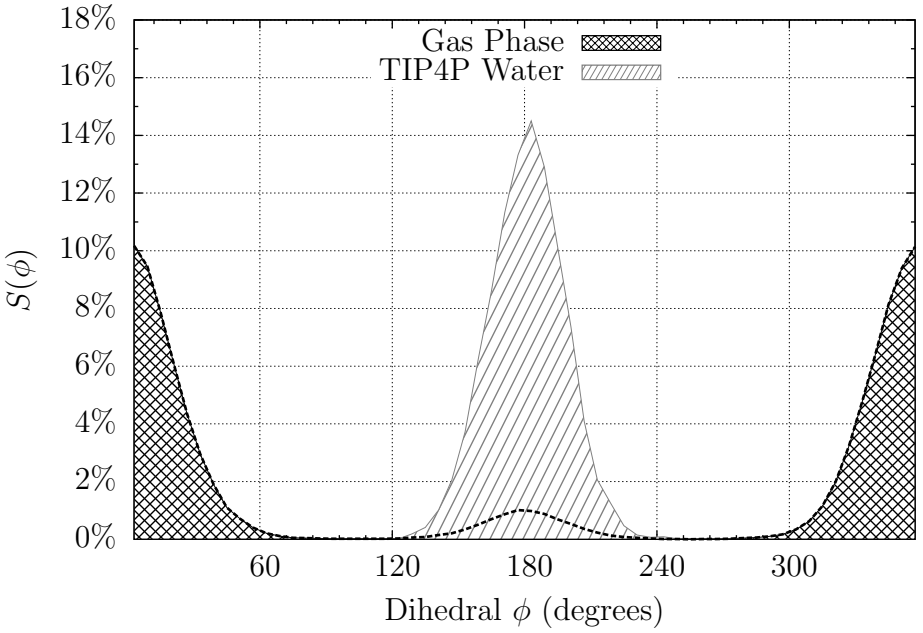


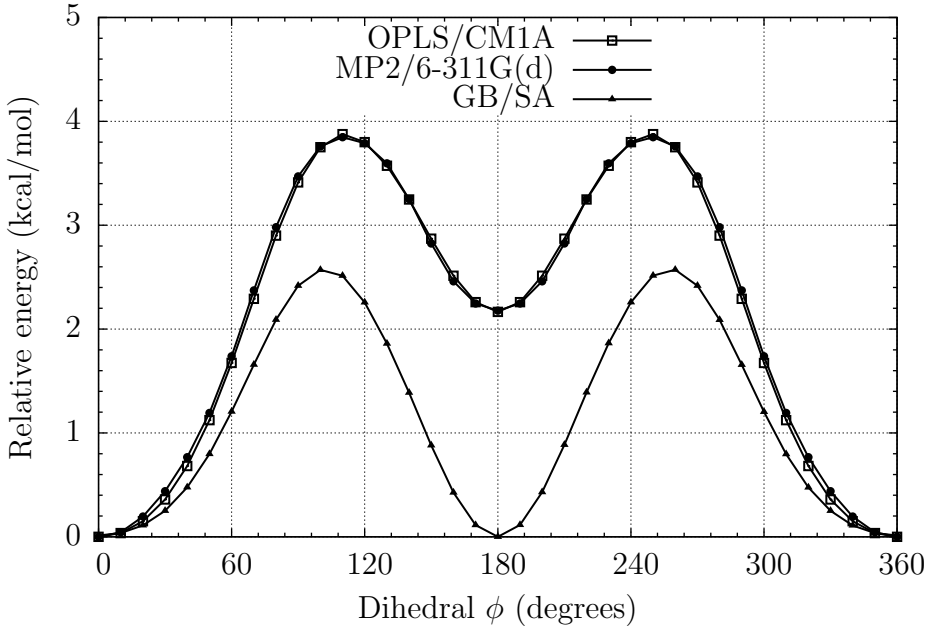
Figure 3.11: Torsion profile and dihedral angle distributions for pyridine-2-thiol.

top. The less polar (1.11 D) global minimum was found at a CT–S–CW–N dihedral angle of 0° (eclipsed), and a more polar (4.0 D) local minimum was found at 180°, at a relative energy of ca. 2.5 kcal/mol and with an intervening rotational energy barrier of ca. 4 kcal/mol. A shift in population was found to occur between the gas phase and aqueous environments, Figure 3.12, bottom. Of note is that the population was not shifted exclusively: whereas in the gas phase, the population was found to be clustered exclusively around the less polar global minimum at 0°, in water a significant portion of the population remained at this angle, with an approximately equal percent population shifting to the more polar local minimum at 180°. This shift is reflected in the GB/SA torsion profile, where the retention of 0° conformers in water can be seen as a result of the modest (ca. 2.6 kcal/mol) energy barrier.

Moving to the furan series, 2-hydroxyfuran was found to exhibit a similar torsion profile and dihedral angle distribution as pyridine-2-thiol. The dihedral angle distribution (Figure 3.13, bottom) shows a significant shift from the less polar (1.4 D) global minimum at 0° (HO–OH–CW–OS) to the more polar (2.8 D) local minimum at 180°. In fact, as can be seen clearly in the GB/SA torsion profile, the global and local minima again flipped, or inverted, in going from the gas phase to an aqueous environment (Figure 3.13, top), with the more polar 180° conformer becoming the new global minimum and the less polar 0° conformer becoming the new local minimum.

Finally, we turn to 2-methoxyfuran, Figure 3.14. This was a somewhat unique case in that the rotational energy barriers were found to have increased, rather than decreased, in going from gas to aqueous environments, Figure 3.14, top. The consequence of this is reflected in the dihedral angle distributions (Figure 3.14, bottom), where the shift occurred from the less polar (1.8 D) local minimum at 60/300° (CT–OS–CW–OS) to the more polar (2.75 D) global minimum at 180°; in the previous four cases detailed here, shifts occurred from the global minimum to the local minima. Thus, owing to an increased rotational energy barrier in water, the global and local minima became more

Torsion profile for 2-(methylthio)pyridine.



Dihedral angle distributions for 2-(methylthio)pyridine.

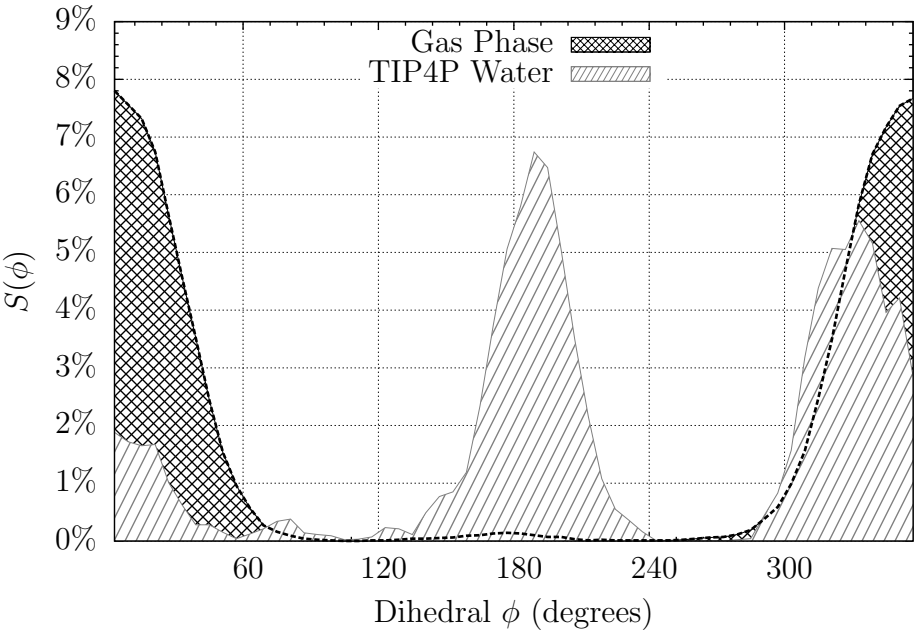
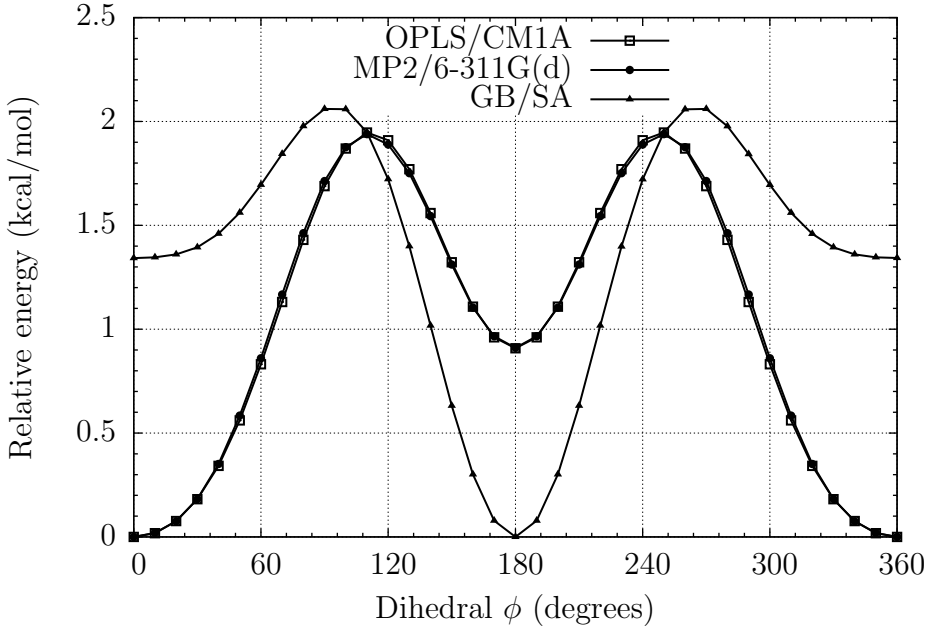


Figure 3.12: Torsion profile and dihedral angle distributions for 2-(methylthio)pyridine.

Torsion profile for 2-hydroxyfuran.



Dihedral angle distributions for 2-hydroxyfuran.

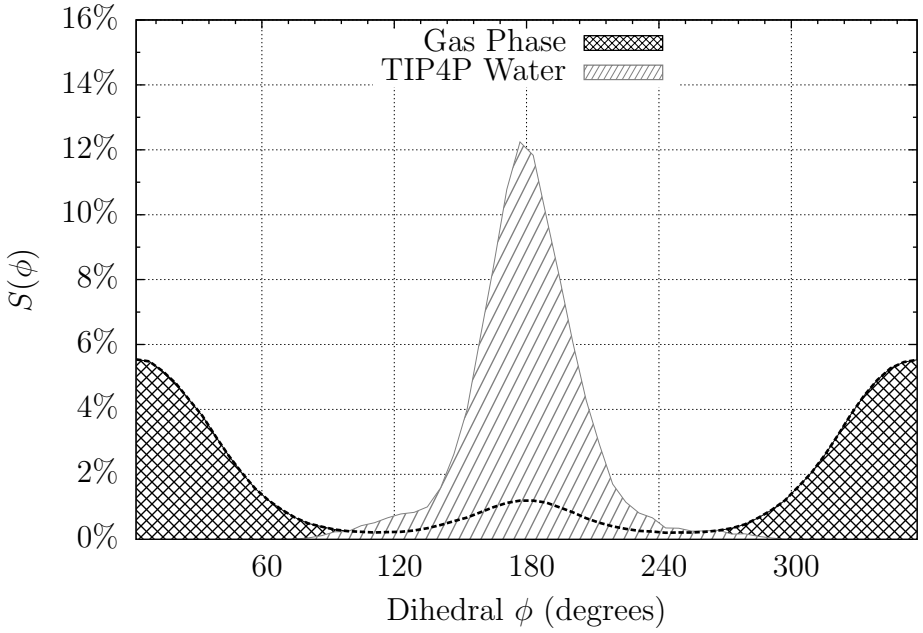
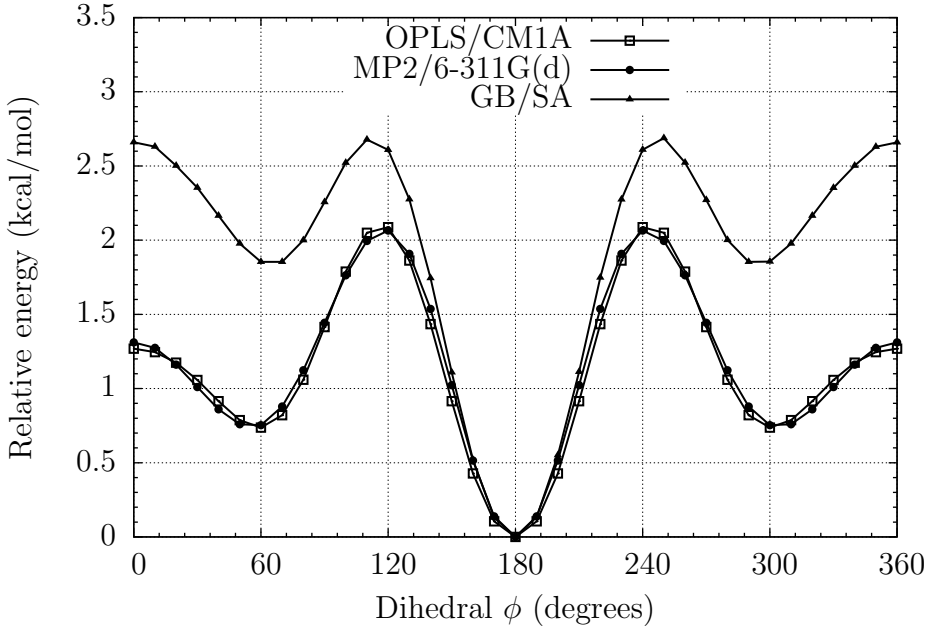


Figure 3.13: Torsion profile and dihedral angle distributions for 2-hydroxyfuran.

Torsion profile for 2-methoxyfuran.



Dihedral angle distributions for 2-methoxyfuran.

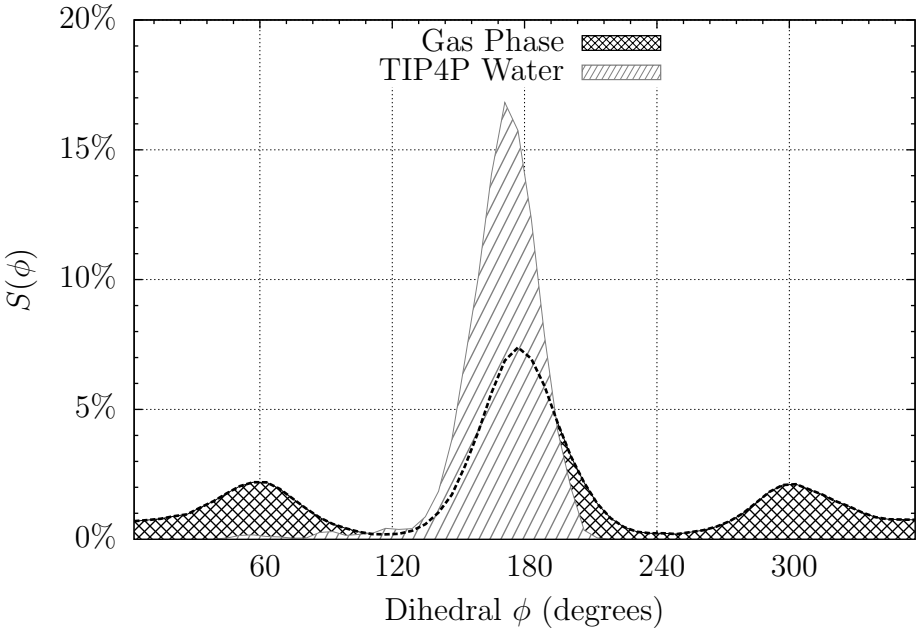


Figure 3.14: Torsion profile and dihedral angle distributions for 2-methoxyfuran.

(rather than less) energetically well separated and the population at the local minimum vanishes; this shift is also aided in part by the associated change in dipole between the two conformers.

3.3 Conclusions.

For 65 prototypical derivatives of benzene, pyridine, furan, thiophene, and pyrrole, containing methyl, ethyl, isopropyl, cyclopropyl, *t*-butyl, vinyl, hydroxy, methoxy, thio, methylthio, amino, *N*-methylamino, and *N*,*N*-dimethylamino substituents, torsion parameters for the OPLS-AA force field were developed and fit to quantum mechanical data. The parameterization yielded well-defined torsion curves that mimicked the quantum mechanically calculated curves to less than 0.5 kcal/mol error and afforded 66.9%–95.8% reduction of error over unparameterized molecular mechanical curves. Gas-phase Monte Carlo dihedral angle distributions mimicked gas-phase torsion profiles and aqueous Monte Carlo dihedral angle distributions mimicked GB/SA torsion profiles. For a small subset of the 65 molecules (OH, SH, and SCH_3 derivatives of 2-pyridine and OH and OCH_3 derivatives of 2-furan), a shift in dihedral angle population was observed in transfer from gas-phase to aqueous environments owing to significantly different torsion profiles between gas and GB/SA. Results are fundamentally important to the development of molecular modeling software and in understanding small-molecule conformational energetics and dynamics: well-predicted dihedral populations for given molecular species can aid computational, medicinal, and organic chemists in making the correct choices in molecular designs to achieve the desired molecular shape in a given environment. The rationale for the choice of heterocycles and derivatives in the context of drug design was given in Chapter 1 of this dissertation and the need for such development was highlighted in Chapter 2; without a doubt, progress in continuum solvation models and force field development will continue to play an ever-increasingly pivotal role in the future of chemical theory, application, and the myriad of interfaces between.

References

1. Heitler, W.; London, F. "Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik." *Physik* **1927**, *44*, 455–473.

2. Pauling, L. "The nature of the chemical bond. Application of results obtained from the quantum mechanics and from a theory of paramagnetic susceptibility to the structure of molecules." *The Journal of the American Chemical Society* **1931**, *53*(4), 1367–1400.

3. Boys, S. F.; Cook, G. B.; Reeves, C. M.; Shavitt, I. "Automatic fundamental calculations of molecular structure." *Nature* **1956**, *178*(4544), 1207–1209.

4. Smith, S. J.; Sutcliffe, B. T. "The development of computational chemistry in the United Kingdom." *Reviews in Computational Chemistry* **2007**, *10*, 271–316.

5. Bernal, J. D.; Fowler, R. H. "A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions." *The Journal of Chemical Physics* **1933**, *1*(8), 515–548.

6. Stillinger, F. H.; Rahman, A. "Improved simulation of liquid water by molecular dynamics." *The Journal of Chemical Physics* **1974**, *60*(4), 1545–1557.

7. Jorgensen, W. L. "Quantum and statistical mechanical studies of liquids. 11. Transferable intermolecular potential functions. Application to liquid methanol including internal rotation." *The Journal of the American Chemical Society* **1981**, *103*(2), 341–345.

8. Jorgensen, W. L. "Revised TIPS for simulations of liquid water and aqueous solutions." *The Journal of Chemical Physics* **1982**, *77*(8), 4156–4163.

9. Jorgensen, W. L.; Madura, J. D. "Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water." *The Journal of the American Chemical Society* **1983**, *105*(6), 1407–1413.

10. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. "The missing term in effective pair potentials." *The Journal of Physical Chemistry* **1987**, *91*(24), 6269–6271.

11. Mahoney, M. W.; Jorgensen, W. L. "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions." *The Journal of Chemical Physics* **2000**, *112*(20), 8910–8922.

12. Nada, H.; van der Eerden, J. P. J. M. "An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H_2O." *The Journal of Chemical Physics* **2003**, *118*(16), 7401–7413.

13. Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. "Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew." *The Journal of Chemical Physics* **2004**, *120*(20), 9665–9678.

14. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. "Gaussian-3 (G3) theory for molecules containing first- and second-row atoms." *The Journal of Chemical Physics* **1998**, *109*(18), 7764–7776.

15. Pople, J. A.; Santry, D. P.; Segal, G. A. "Approximate self-consistent molecular orbital theory. I. Invariant procedures." *The Journal of Chemical Physics* **1965**, *43*(10), S129–S135.

16. Pople, J. A.; Segal, G. A. "Approximate self-consistent molecular orbital theory. II. Calculations with complete neglect of differential overlap." *The Journal of Chemical Physics* **1965**, *43*(10), S136–S151.

17. Santry, D. P.; Segal, G. A. "Approximate self-consistent molecular orbital theory. IV. Calculations on molecules including the elements sodium through chlorine." *The Journal of Chemical Physics* **1967**, *47*(1), 158–174.

18. Pople, J. A.; Beveridge, D. L.; Dobosh, P. A. "Approximate self-consistent molecular-orbital theory. V. Intermediate neglect of differential overlap." *The Journal of Chemical Physics* **1967**, *47*(6), 2026–2033.

19. Allinger, N. L. "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V_1 and V_2 torsional terms." *The Journal of the American Chemical Society* **1977**, *99*(25), 8127–8134.

20. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. "Development and use of quantum mechanical molecular models. 76. AM1: a new general-purpose quantum mechanical molecular model." *The Journal of the American Chemical Society* **1985**, *107*(13), 3902–3909.

21. Jorgensen, W. L.; Tirado-Rives, J. "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin." *The Journal of the American Chemical Society* **1988**, *110*(6), 1657–1666.

22. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids." *The Journal of the American Chemical Society* **1996**, *118*(45), 11225–11236.

23. Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. "Prediction of the water content

in protein binding sites." *The Journal of Physical Chemistry B* **2009**, *113*(40), 13337–13346.

24. Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. "Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization." *The Journal of the American Chemical Society* **2009**, *131*(42), 15403–15411.

25. Luccarelli, J.; Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. "Effects of water placement on predictions of binding affinities for p38α MAP kinase inhibitors." *The Journal of Chemical Theory and Computation* **2010**, *6*(12), 3850–3856.

26. Orozco, M.; Luque, F. J. "Theoretical methods for the description of the solvent effect in biomolecular systems." *Chemical Reviews* **2000**, *100*(11), 4187–4226.

27. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. "Semianalytical treatment of solvation for molecular mechanics and dynamics." *The Journal of the American Chemical Society* **1990**, *112*(16), 6127–6129.

28. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. "The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii." *The Journal of Physical Chemistry A* **1997**, *101*(16), 3005–3014.

29. Hermann, R. B. "Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area." *The Journal of Physical Chemistry* **1972**, *76*(19), 2754–2759.

30. Floris, F.; Tomasi, J. "Evaluation of the dispersion contribution to the solvation energy. A simple computational model in the continuum approximation." *The Journal of Computational Chemistry* **1989**, *10*(5), 616–627.

31. Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. "The Gaussian

generalized Born model: application to small molecules." *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913–4922.

32. Jayaram, B.; Liu, Y.; Beveridge, D. L. "A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts." *The Journal of Chemical Physics* **1998**, *109*(4), 1465–1471.

33. Jorgensen, W. L.; Tirado-Rives, J. "Molecular modeling of organic and biomolecular systems using BOSS and MCPRO." *The Journal of Computational Chemistry* **2005**, *26*(16), 1689–1700.

34. Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. "Free energies of hydration from a generalized Born model and an all-atom force field." *The Journal of Physical Chemistry B* **2004**, *108*(41), 16264–16270.

35. Bashford, D.; Case, D. A. "Generalized Born models of macromolecular solvation effects." *Annual Review of Physical Chemistry* **2000**, *51*(1), 129–152.

36. Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. "RNA simulations: probing hairpin unfolding and the dynamics of a GNRA tetraloop." *The Journal of Molecular Biology* **2002**, *317*(4), 493–506.

37. Sorin, E. J.; Rhee, Y. M.; Nakatani, B. J.; Pande, V. S. "Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations." *Biophysical Journal* **2003**, *85*(2), 790–803.

38. Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. "Free energy surfaces of β-hairpin and α-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model." *Proteins: Structure, Function, and Bioinformatics* **2004**, *56*(2), 310–321.

39. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics* **1953**, *21*(6), 1087–1092.

40. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. "FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function." *The Journal of Computational Chemistry* **2003**, *24*(13), 1637–1656.

41. Ulmschneider, J. P.; Jorgensen, W. L. "Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias." *The Journal of Chemical Physics* **2003**, *118*(9), 4261–4271.

42. Ulmschneider, J. P.; Jorgensen, W. L. "Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation." *The Journal of the American Chemical Society* **2004**, *126*(6), 1849–1857.

43. Simonson, T.; Carlsson, J.; Case, D. A. "Proton binding to proteins: pK_{a} calculations with explicit and implicit solvent models." *The Journal of the American Chemical Society* **2004**, *126*(13), 4167–4180.

44. Henchman, R. H.; Kilburn, J. D.; Turner, D. L.; Essex, J. W. "Conformational and enantioselectivity in host–guest chemistry: the selective binding of *cis* amides examined by free energy calculations." *The Journal of Physical Chemistry B* **2004**, *108*(45), 17571–17582.

45. Gallicchio, E.; Zhang, L. Y.; Levy, R. M. "The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimators." *The Journal of Computational Chemistry* **2002**, *23*(5), 517–529.

46. Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. "Solvent models for protein–ligand binding: comparison of implicit solvent poisson and surface generalized Born

models with explicit solvent simulations." *The Journal of Computational Chemistry* **2001**, *22*(6), 591–607.

47. Michel, J.; Verdonk, M. L.; Essex, J. W. "Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization." *The Journal of Medicinal Chemistry* **2006**, *49*(25), 7427–7439.

48. Guvench, O.; Weiser, J.; Shenkin, P.; Kolossváry, I.; Still, W. C. "Application of the frozen atom approximation to the GB/SA continuum model for solvation free energy." *The Journal of Computational Chemistry* **2002**, *23*(2), 214–221.

49. Michel, J.; Taylor, R. D.; Essex, J. W. "Efficient generalized Born models for Monte Carlo simulations." *The Journal of Chemical Theory and Computation* **2006**, *2*(3), 732–739.

50. Gelb, L. D. "Monte Carlo simulations using sampling from an approximate potential." *The Journal of Chemical Physics* **2003**, *118*(17), 7747–7750.

51. Felts, A. K.; Gallicchio, E.; Chekmarev, D.; Paris, K. A.; Friesner, R. A.; Levy, R. M. "Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling." *The Journal of Chemical Theory and Computation* **2008**, *4*(5), 855–868.

52. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. "Pairwise solute descreening of solute charges from a dielectric medium." *Chemical Physics Letters* **1995**, *246*(1), 122–129.

53. Zwanzig, R. W. "High-temperature equation of state by a perturbation method. I. Nonpolar gases." *The Journal of Chemical Physics* **1954**, *22*(8), 1420–1426.

54. Jorgensen, W. L.; Thomas, L. L. "Perspective on free-energy perturbation calculations for chemical equilibria." *The Journal of Chemical Theory and Computation* **2008**, *4*(6), 869–876.

55. Jorgensen, W. L. "The many roles of computation in drug discovery." *Science* **2004**, *303*(5665), 1813–1818.

56. Barreiro, G.; Guimarães, C. R. W.; Tubert-Brohman, I.; Lyons, T. M.; Tirado-Rives, J.; Jorgensen, W. L. "Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring." *The Journal of Chemical Information and Modeling* **2007**, *47*(6), 2416–2428.

57. Barreiro, G.; Kim, J. T.; Guimarães, C. R. W.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. "From docking false-positive to active anti-HIV agent." *The Journal of Medicinal Chemistry* **2007**, *50*(22), 5324–5329.

58. Jorgensen, W. L.; Tirado-Rives, J. *MCPRO*, Yale University, New Haven, CT **2006**.

59. Zeevaart, J. G.; Wang, L.; Thakur, V. V.; Leung, C. S.; Tirado-Rives, J.; Bailey, C. M.; Domaoal, R. A.; Anderson, K. S.; Jorgensen, W. L. "Optimization of azoles as anti-human immunodeficiency virus agents guided by free-energy calculations." *The Journal of the American Chemical Society* **2008**, *130*(29), 9492–9499.

60. Leung, C. S.; Zeevaart, J. G.; Domaoal, R. A.; Bollini, M.; Thakur, V. V.; Spasov, K. A.; Anderson, K. S.; Jorgensen, W. L. "Eastern extension of azoles as non-nucleoside inhibitors of HIV-1 reverse transcriptase; cyano group alternatives." *Bioorganic & Medicinal Chemistry Letters* **2010**, *20*(8), 2485–2488.

61. Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. "Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis." *The Journal of Chemical Information and Modeling* **2008**, *48*(1), 1–24.

62. Jabs, A.; Weiss, M. S.; Hilgenfeld, R. "Non-proline *cis* peptide bonds in proteins." *The Journal of Molecular Biology* **1999**, *286*(1), 291–304.

63. Wiberg, K. B.; Laidig, K. E. "Barriers to rotation adjacent to double bonds. 3. The carbon-oxygen barrier in formic acid, methyl formate, acetic acid, and methyl acetate. The origin of ester and amide resonance." *The Journal of the American Chemical Society* **1987**, *109*(20), 5935–5943.

64. Jorgensen, W. L.; Gao, J. "*Cis–trans* energy difference for the peptide bond in the gas phase and in aqueous solution." *The Journal of the American Chemical Society* **1988**, *110*(13), 4212–4216.

65. Remko, M.; Scheiner, S. "The geometry and internal rotational barrier of carbamic acid and several derivatives." *The Journal of Molecular Structure: THEOCHEM* **1988**, *180*, 175–188.

66. Glaser, R.; Streitwieser, A. "Configurational and conformational preferences in oximes and oxime carbanions. Ab initio study of the *syn* effect in reactions of oxyimine enolate equivalents." *The Journal of the American Chemical Society* **1989**, *111*(19), 7340–7348.

67. Stang, P. J.; Kitamura, T.; Karni, M.; Apeloig, Y.; Arif, A. M. "A single crystal molecular structure determination and theoretical calculations on alkynyl carboxylate esters." *The Journal of the American Chemical Society* **1990**, *112*(1), 374–381.

68. Wiberg, K. B.; Wong, M. W. "Solvent effects. 4. Effect of solvent on the E/Z energy difference for methyl formate and methyl acetate." *The Journal of the American Chemical Society* **1993**, *115*(3), 1078–1084.

69. Evanseck, J. D.; Houk, K. N.; Briggs, J. M.; Jorgensen, W. L. "Quantification of solvent effects on the acidities of Z and E esters from fluid simulations." *The Journal of the American Chemical Society* **1994**, *116*(23), 10630–10638.

70. Deerfield, D. W.; Pedersen, L. G. "An ab initio quantum mechanical study of

thioesters." *The Journal of Molecular Structure: THEOCHEM* **1995**, *358*(1), 99–106.

71. Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. "Ab initio calculations on small molecule analogs of polycarbonates." *The Journal of Physical Chemistry* **1995**, *99*(16), 5873–5882.

72. Strassner, T. "Ab initio and molecular mechanics calculations of various substituted ureas: rotational barriers and a new parametrization for ureas." *Molecular modeling annual* **1996**, *2*(9), 217–226.

73. Murphy, R. B.; Pollard, W. T.; Friesner, R. A. "Pseudospectral localized generalized Møller–Plesset methods with a generalized valence bond reference wave function: Theory and calculation of conformational energies." *The Journal of Chemical Physics* **1997**, *106*(12), 5073–5084.

74. Chambers, C. C.; Archibong, E. F.; Jabalameli, A.; Sullivan, R. H.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. "Quantum mechanical and ^{13}C dynamic NMR study of 1,3-dimethylthiourea conformational isomerizations." *The Journal of Molecular Structure: THEOCHEM* **1998**, *425*(1), 61–68.

75. Császár, A. G.; Allen, W. D.; Schaefer, H. F. "In pursuit of the ab initio limit for conformational energy prototypes." *The Journal of Chemical Physics* **1998**, *108*(23), 9751–9764.

76. Villani, V.; Alagona, G.; Ghio, C. "Ab initio studies on *N*-methylacetamide." *Molecular Engineering* **1998**, *8*(2), 135–153.

77. Kang, Y. K. "Ab initio MO and density functional studies on *trans* and *cis* conformers of *N*-methylacetamide." *The Journal of Molecular Structure: THEOCHEM* **2001**, *546*(1), 183–193.

78. Senent, M. L. "Ab initio determination of the torsional spectra of acetic acid." *Molecular Physics* **2001**, *99*(15), 1311–1321.

79. Kobychev, V. B.; Vitkovskaya, N. M.; Pavlova, N. V.; Schmidt, E. Y.; Trofimov, B. A. "Theoretical analysis and experimental study of the spatial structure and isomerism of acetone azine and its cyclization to 3,5,5-trimethyl-4,5-dihydro-1*H*-pyrazole." *The Journal of Structural Chemistry* **2004**, *45*(5), 748–755.

80. Zhong, H.; Stewart, E. L.; Kontoyianni, M.; Bowen, J. P. "Ab initio and DFT conformational studies of propanal, 2-butanone, and analogous imines and enamines." *The Journal of Chemical Theory and Computation* **2005**, *1*(2), 230–238.

81. Bryantsev, V. S.; Firman, T. K.; Hay, B. P. "Conformational analysis and rotational barriers of alkyl- and phenyl-substituted urea derivatives." *The Journal of Physical Chemistry A* **2005**, *109*(5), 832–842.

82. Bryantsev, V. S.; Hay, B. P. "Conformational preferences and internal rotation in alkyl- and phenyl-substituted thiourea derivatives." *The Journal of Physical Chemistry A* **2006**, *110*(14), 4678–4688.

83. Mantz, Y. A.; Branduardi, D.; Bussi, G.; Parrinello, M. "Ensemble of transition state structures for the *cis–trans* isomerization of *N*-methylacetamide." *The Journal of Physical Chemistry B* **2009**, *113*(37), 12521–12529.

84. Klebe, G. "Virtual ligand screening: strategies, perspectives and limitations." *Drug Discovery Today* **2006**, *11*(13), 580–594.

85. Jorgensen, W. L. "Efficient drug lead discovery and optimization." *Accounts of Chemical Research* **2009**, *42*(6), 724–733.

86. Nichols, S. E.; Domaoal, R. A.; Thakur, V. V.; Tirado-Rives, J.; Anderson, K. S.; Jorgensen, W. L. "Discovery of wild-type and Y181C mutant non-nucleoside HIV-

1 reverse transcriptase inhibitors using virtual screening with multiple protein structures." *The Journal of Chemical Information and Modeling* **2009**, *49*(5), 1272–1279.

87. Ren, J.; Diprose, J.; Warren, J.; Esnouf, R. M.; Bird, L. E.; Ikemizu, S.; Slater, M.; Milton, J.; Balzarini, J.; Stuart, D. I.; Stammers, D. K. "Phenylethylthiazolylthiourea (PETT) Non-nucleoside Inhibitors of HIV-1 and HIV-2 Reverse Transcriptases: Structural and Biochemical Analyses." *The Journal of Biological Chemistry* **2000**, *275*(8), 5633–5639.

88. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. "Prediction of protein–ligand interactions. Docking and scoring: successes and gaps." *The Journal of Medicinal Chemistry* **2006**, *49*(20), 5851–5855.

89. Ponder, J. W.; Case, D. A. "Force fields for protein simulations." *Advances in Protein Chemistry* **2003**, *66*, 27–85.

90. Jorgensen, W. L.; Tirado-Rives, J. "Potential energy functions for atomic-level simulations of water and organic and biomolecular systems." *Proceedings of the National Academy of Sciences* **2005**, *102*(19), 6665–6670.

91. Gaussian 03, Revision C.02. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.;

Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian, Inc.*, Wallingford, CT **2004**.

92. Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. "Gaussian-3 theory using density functional geometries and zero-point energies." *The Journal of Chemical Physics* **1999**, *110*(16), 7650–7657.

93. Maçôas, E. M. S.; Khriachtchev, L.; Pettersson, M.; Fausto, R.; Räsänen, M. "Rotational isomerism in acetic acid: the first experimental observation of the high-energy conformer." *The Journal of the American Chemical Society* **2003**, *125*(52), 16188–16189.

94. Pawar, D. M.; Khalil, A. A.; Hooks, D. R.; Collins, K.; Elliott, T.; Stafford, J.; Smith, L.; Noe, E. A. "*E* and *Z* conformations of esters, thiol esters, and amides." *The Journal of the American Chemical Society* **1998**, *120*(9), 2108–2112.

95. Huisgen, R.; Ott, H. "Die konfiguration der carbonestergruppe und die sondereigenschaften der lactone." *Tetrahedron* **1959**, *6*(3), 253–267.

96. Schweizer, W. B.; Dunitz, J. D. "Structural characteristics of the carboxylic ester group." *Helvetica Chimica Acta* **1982**, *65*(5), 1547–1554.

97. Nobes, R. H.; Radom, L.; Allinger, N. L. "Equilibrium conformations of higher-energy rotational isomers of vinyl alcohol and methyl vinyl ether." *The Journal of Molecular Structure: THEOCHEM* **1981**, *85*(1), 185–194.

98. Yamabe, S.; Tsuchida, N.; Yamazaki, S. "Is the Beckmann rearrangement a concerted

or stepwise reaction? A computational study." *The Journal of Organic Chemistry* **2005**, *70*(26), 10638–10644.

99. Gao, J.; Pavelites, J. J. "Aqueous basicity of the carboxylate lone pairs and the carbon–oxygen barrier in acetic acid: a combined quantum and statistical mechanical study." *The Journal of the American Chemical Society* **1992**, *114*(5), 1912–1914.

100. Sato, H.; Hirata, F. "The *syn*/*anti* conformational equilibrium of acetic acid in water studied by the RISM-SCF/MCSCF method." *The Journal of Molecular Structure: THEOCHEM* **1999**, *461-462*(2), 113–120.

101. Li, Y.; Houk, K. N. "Theoretical assessments of the basicity and nucleophilicity of carboxylate *syn* and *anti* lone pairs." *The Journal of the American Chemical Society* **1989**, *111*(12), 4505–4507.

102. Rebek Jr., J. "Molecular recognition with model systems." *Angewandte Chemie International Edition in English* **1990**, *29*(3), 245–255.

103. Kroeger Smith, M. B.; Hose, B. M.; Hawkins, A.; Lipchock, J.; Farnsworth, D. W.; Rizzo, R. C.; Tirado-Rives, J.; Arnold, E.; Zhang, W.; Hughes, S. H.; Jorgensen, W. L.; Michejda, C. J.; Smith, R. H. "Molecular modeling calculations of HIV-1 reverse transcriptase non-nucleoside inhibitors: correlation of binding energy with biological activity for novel 2-aryl-substituted benzimidazole analogues." *The Journal of Medicinal Chemistry* **2003**, *46*(10), 1940–1947.

104. Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. "Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase." *Bioorganic & Medicinal Chemistry Letters* **2006**, *16*(3), 663–667.

105. Thakur, V. V.; Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L. "Optimization of pyrimidinyl- and triazinyl-

amines as non-nucleoside inhibitors of HIV-1 reverse transcriptase." *Bioorganic & Medicinal Chemistry Letters* **2006**, *16*(21), 5664–5667.

106. Venkatachalam, T. K.; Uckun, F. M. "Regiospecific synthesis of 5-halo-substituted thiophene pyridyl thiourea compounds as non-nucleoside inhibitors of HIV-1 reverse transcriptase." *Synthetic Communications* **2004**, *34*(13), 2451–2461.

107. Antonucci, T.; Warmus, J. S.; Hodges, J. C.; Nickell, D. G. "Characterization of the antiviral activity of highly substituted pyrroles: a novel class of non-nucleoside HIV-1 reverse transcriptase inhibitor." *Antiviral Chemistry and Chemotherapy* **1995**, *6*(2), 98–108.

108. Terhorst, J. P.; Jorgensen, W. L. "E/Z energetics for molecular modeling and design." *The Journal of Chemical Theory and Computation* **2010**, *6*(9), 2762–2769.

109. Kollman, P. "Free energy calculations: applications to chemical and biochemical phenomena." *Chemical Reviews* **1993**, *93*(7), 2395–2417.

110. Briggs, J. M.; Matsui, T.; Jorgensen, W. L. "Monte Carlo simulations of liquid alkyl ethers with the OPLS potential functions." *The Journal of Computational Chemistry* **1990**, *11*(8), 958–971.

Appendix

A1. New torsion parameters for derivatives of benzene and pyridine.

Benzenes					Pyridines				
Dihedral	V_1	V_2	V_3	V_4	Dihedral	V_1	V_2	V_3	V_4
CT–OS–CA–CA	0	3.05	0	0.3	HC–CT–CA–NC	0.73	0.03	0.39	-0.01
HC–CT–CA–CA	0	0	0.35	0	HC–C=–CA–NC	-0.38	4.46	0.55	-0.12
CT–CT–CA–CA	0	0	-0.25	0	CM–C=–CA–NC	0	0	0	0
HO–OH–CA–CA	0	2.4	0	0	HO–OH–CA–NC	-0.3	4.4	0	0
C?–CA–SH–HS	0	0	0	-0.07	HS–SH–CA–NC	0.7	3.1	0.5	0
CA–CA–SH–HS	0	0	0	0	CY–CY–CA–NC	0	0	0	0
CA–CA–NT–H	0	2.2	0	0	H–N–CA–NC	0.15	3.44	-1.39	1.4
HC–C=–CA–CA	0	-0.05	0	0	CT–N–CA–NC	0	2	-1	0
CM–C=–CA–CA	0	3.2	0	0	HC–CY–CA–NC	1.24	3.27	0.43	0.17
CT–S–CA–CA	0	0.29	0	0.05	CT–OS–CA–NC	-0.5	4.65	0.5	0
CY–CY–CA–CA	0	1.315	0	0	CT–S–CA–NC	0.85	3.2	0.9	0
HC–CY–CA–CA	0	0.3	0	0	CT–CT–CA–NC	0	0	0.1	0

A2. New torsion parameters for derivatives of pyrrole and thiophene.

Pyrroles					Thiophenes				
Dihedral	V_1	V_2	V_3	V_4	Dihedral	V_1	V_2	V_3	V_4
CT–CT–CW–NA	1.91	-1.82	0.2	-0.9	HC–CT–CW–S	0	0	0.16	0
HC–CT–CW–NA	-0.08	-1.51	0.225	-0.63	CT–CT–CW–S	-1.44	-0.13	-0.03	-0.13
HC–CY–CW–NA	-2.19	-0.21	0.28	0.09	HC–CY–CW–S	2.38	-1.2	0.37	0
CY–CY–CW–NA	0	0	0	0	CM–C=–CW–S	-2	5.7	0	-0.3
H–N2–CW–CS	0	0	0	0	HC–C=–CW–S	0	0	0	0
H–N2–CW–NA	3.29	0.08	0	-0.08	HO–OH–CW–S	3.22	1.05	0.46	-0.09
CT–N2–CW–CS	0	0	0	0	CT–OS–CW–S	4.9	3.5	1.4	0.5
CT–N2–CW–NA	6	1	0	0	HS–SH–CW–S	0.9	-2.93	0.32	-0.1
HO–OH–CW–NA	0.21	0.64	0.2	-0.05	CT–S–CW–S	0.97	-2.19	0.58	0.26
CT–OS–CW–NA	5.79	2.44	1.27	0.81	H–N–CW–S	-2.91	0.64	-0.22	0.44
HS–SH–CW–NA	0.2	-4.54	0.17	-0.25	CT–N–CW–S	-3.75	1.99	0.34	-0.29
CT–S–CW–NA	3.19	-4.35	0.65	-0.05					
CM–C=–CW–NA	1.6	6.76	-0.29	-0.26					
HC–C=–CW–NA	0	0	0	0					

A3. New torsion parameters for derivatives of furan.

Furans									
Dihedral	V_1	V_2	V_3	V_4	Dihedral	V_1	V_2	V_3	V_4
HC–CT–CW–OS	0	0	0.29	0	HO–OH–CW–OS	1.35	1.55	0.5	0
HC–CT–CW–CS	0	0	0	0	HO–OH–CW–CS	0	0	0	0
CT–CT–CW–CS	0	0	0	0	HS–SH–CW–CS	0	0	0	0
CT–CT–CW–OS	-0.075	0	0.525	-0.1	HS–SH–CW–OS	0.6	-2.2	0.8	0
HC–CY–CW–CS	0	0	0	0	CT–OS–CW–CS	0	0	0	0
CY–CY–CW–OS	0	0	0	0	CT–OS–CW–OS	4.9	3.4	2.1	0.55
CY–CY–CW–CS	0	0	0	0	CT–S–CW–CS	0	0	0	0
CM–C=–CW–CS	0	0	0	0	CT–S–CW–OS	1.5	-1.7	1.5	0.3
HC–CY–CW–OS	0.81	1	1.1	0.15	H–N–CW–CS	0	0	0	0
CM–C=–CW–OS	-0.9	7.15	-0.9	0	OS–CW–N–N	-0.4	1.4	2	0
HC–C=–CW–CS	0	0	0	0	CT–N–CW–OS	0	2.6	3.6	0
HC–C=–CW–OS	0	0	0	0	CT–N–CW–CS	0	0	0	0