Computational Investigation of Nucleic Acids

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Adam Thomas Moser

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor Of Philosophy

July, 2009

© Adam Thomas Moser 2009
ALL RIGHTS RESERVED

Acknowledgements

I would like to give my deepest thanks to the following. The National Defense Science and Engineering Graduate (NDSEG) fellowship for funding. The Minnesota Supercomputing Institute for Advanced Computational Research for supplying both computational resources and as well as technical assistance. To the members of the York lab, in particular Dr. Kevin Range and Dr. Tim Giese. To my collaborators at the University of Minnesota Masonic Cancer Center, Rebecca Guza, Uthpala Seneviratne, and Dr. Natalia Tretyakova. Susan Ballinger, who started me on this path. Professor Scott Feller at Wabash College, who is my role model. To my adviser Professor Darrin York, who has been open to my ideas and understanding of my goals. Lastly, to my family, whose constant support has helped keep me going.

Dedication

This is dedicated to all my future students. You better be worth it.

ABSTRACT

In this work, various computational chemistry models are applied to problems of biochemical interest, with emphasis on nucleic acids. First various density functionals and multilevel methods are benchmarked against experimental proton affinities and gas-phase basicities. Then prediction of biologically relevant values of nucleic acids, amino acids, RNA sugar, and phosphates are made. In applied work, density functional theory is employed to help elucidate topics in lesion formation in nucleic acids. In particular the role of C5 methyl cytosine substitution is investigated through the use of various analogues and explaining NMR spectra of specific adenine lesions formed by 1,2,3,4-diepoxybutane. Finally, two works related to parameterization are given. The first is CHARMM molecular mechanical force field parameter development for the reactive intermediates of native and thio-substituted ribozymes. This work provides modifications necessary to reproduce structural aspects of transition state structures during phosphate transesterification. The second is an investigation into the appropriate solvation free energy for phosphoric acid and its anions. This includes both a review of the currently used and available data as well as a benchmark of various computational solvation models.

Contents

Acknowledgements **i**

Dedication **ii**

Abstract **iii**

List of Tables **vii**

List of Figures **x**

Preface **xii**

1 Modeling biological systems through computational chemistry **1**

1.1 Computational chemistry 1

1.2 Challenges in modeling biomolecules 2

1.3 Electronic structure 3

1.3.1 Hartree-Fock and Semiempirical 5

1.3.2 Density functional theory 6

1.3.3 Multilevel methods 8

1.4 Molecular mechanics 9

1.5 Solvation 10

1.6 Multiscale modeling 12

1.7 Preview 14

2 Proton affinity and gas-phase basicity of biocatalytic molecules **16**

2.1 Introduction 16

2.2 Methods . . . 18
2.2.1 Electronic structure calculations. . . . 18
2.2.2 Calculation of proton affinities and gas-phase basicities. . . . 20
2.3 Results and Discussion . . . 22
2.3.1 Experimental Comparison . . . 23
2.3.2 Amino Acids . . . 28
2.3.3 Nucleic Acid Bases . . . 32
2.3.4 Ribose . . . 34
2.3.5 Phosphates . . . 36
2.4 Conclusion . . . 40
2.5 Supporting Information . . . 40

3 Influence of C5 cytosine substitution in base pairs with guanine 52
3.1 Introduction . . . 52
3.2 Methods . . . 54
3.3 Results . . . 58
3.3.1 Cytosine Base . . . 59
3.3.2 CG Base Pair . . . 64
3.3.3 Protonated CG Base Pair . . . 66
3.4 Discussion . . . 67
3.5 Conclusion . . . 72

4 Exocycle lesions of adenine 74
4.1 Introduction . . . 74
4.2 Computational Methods . . . 76
4.3 Results and Discussion . . . 76
4.3.1 Structural determination of N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine . . . 76
4.4 Conclusion . . . 81

5 Parameterization 84
5.1 CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes . . . 84

5.1.1 Background 85
5.1.2 Methods 87
5.1.3 Results and Discussion 93
5.1.4 Conclusion 107
5.1.5 Additional Figures 108
5.2 Solvation of phosphoric acid 112
5.2.1 Introduction 112
5.2.2 Methods 113
5.2.3 Results and Discussion 115
5.2.4 Conclusion 121
5.2.5 Acknowledgment 121

References **122**

List of Tables

2.1 Proton affinity error analysis. 24
2.2 Gas-phase basicity error analysis. 25
2.3 Proton affinity error analysis for bond energy correction (BEC) and linear regression model (LRM). 26
2.4 Proton affinity and gas-phase basicity for amino acid model compounds. 30
2.5 Proton Affinities and gas-phase basicities errors (calculated - experimental value) for DNA and RNA bases. 34
2.6 Proton Affinities and gas-phase basicities errors (calculated - experimental value) for RNA like sugar molecules. 35
2.7 Predicted proton affinities for metaphosphate, phosphate, cyclic and acyclic phosphate, and acyclic and cyclic phosphorane compounds of biological interest. 38
2.8 Coefficients for bond energy correction and linear regression model for proton affinities and gas-phase basicities 41
2.9 Gas-phase basicity error analysis for bond energy correction (BEC) and linear regression model (LRM). 42
2.10 Predicted proton affinities for DNA and RNA bases in keto/amino tautomeric form. 43
2.11 Predicted gas-phase basicity for DNA and RNA bases in keto/amino tautomeric form. 44
2.12 Enthalpy (top) and free energy (bottom) of tautomerization for DNA and RNA basepairs. All quantities are in kcal/mol. 45
2.13 Predicted proton affinities for DNA and RNA bases in enol/imino tautomeric form. 46

2.14 Predicted gas-phase basicity for DNA and RNA bases in enol/imino tautomeric form. 47
2.15 Predicted proton affinities for metaphosphate, phosphate, and cyclic phosphate compounds of biological interest. 48
2.16 Predicted proton affinities for phosphorane compounds of biological interest. 49
2.17 Predicted gas-phase basicities for metaphosphate, phosphate, and cyclic phosphate compounds of biological interest. 50
2.18 Predicted gas-phase basicities for phosphorane compounds of biological interest. 51
3.1 Geometric data for cytosine analogues 60
3.2 Proton affinity, gas phase basicity, and pK_a of various cytosine positions. 62
3.3 GC base pair hydrogen bond geometry. 64
3.4 GC base pair binding energies. 65
3.5 Protonated GC base pair hydrogen bonding geometry 66
3.6 Proton affinity, gas phase basicity, and pK_a of protonated GC base pair. 68
5.1 Mg^{2+} and OH^- parameter fitting results 96
5.2 CHelpG charge fitting for deprotonated ribose phosphate 97
5.3 RNA Lennard-Jones parameters 99
5.4 Geometry fitting results of deprotonated ribose phosphate 99
5.5 CHelpG charge fitting for ribose phosphorane 101
5.6 Geometry fitting results for ribose phosphorane parameterization 102
5.7 Geometry fitting results for 2',3'-cyclic phosphate parameterization . . . 103
5.8 Geometry and binding energy results for phosphate thio-substitution parameterization 105
5.9 Dihedral parameters of non-bridging oxygen and sulfur for thio-substituted phosphate 106
5.10 Geometry fitting results for thio-substituted ribose phosphorane parameterization 106
5.11 Calculated and experimental proton affinities (PA) and gas-phase basicities (GPB) of water, phosphoric acid, dihydrogen phosphate, and hydrogen phosphate. 115

5.12 Experimental literature values for the free energy of solvation (kcal/mol) for phosphoric acid species. 116
5.13 Solvation free energies (kcal/mol) based on gas-phase and solution phase geometry optimizations using various implicit solvation models. 118
5.14 Experimental pK_a values (Ref 342) and W1 calculated gas-phase basicities (See Table 5.11) for phosphoric acid, dihydrogen phosphate, and hydrogen phosphate for different standard states. 118
5.15 Solvation free energies differences (kcal/mol) from structures optimized in the gas-phase and solution. 119
5.16 Partition coefficients for transfer of phosphorus derivatives from water to nonpolar environments (chloroform and vapor) at 20 °C, ionic strength 0.30. 120

List of Figures

1.1 Charge distribution (ρ_0) in a continuum dielectric (ϵ). 12
1.2 Schematic flow of various computational methods and experimental data used in the development of multiscale models. 13
2.1 A commonly employed thermodynamic cycle for calculating pK_{a} values. 17
2.2 Histogram of proton affinity error analysis of CBS and QCRNA using BEC and LRM models 29
2.3 Calculated gas phase basicities of amino acid model compounds versus pK_{a} values. 31
2.4 DNA/RNA basepair protonation points with pK_{a} values. 33
2.5 Nomenclature convention for ligand designations in metaphosphate, acyclic and cyclic phosphate and phosphorane compounds of biological interest. 37
3.1 General steps involved in BPDE reaction with DNA containing $^{\mathrm{Me}}$C . . 53
3.2 Cytosine 5 position analogs 55
3.3 Thermodynamic cycle for pK_{a} prediction. 57
3.4 Watson-Crick GC base pair with guanine N^2 protonation. 57
3.5 Gas phase basicity and hydrogen bonding. 71
4.1 dA lesions from 1,2,3,4-diepoxybutane attack. 75
4.2 N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine. Proton α (red), β (pink), γ (blue), and δ (green). 77
4.3 Axial (top) and equatorial (bottom) conformations of R,R N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine with top and side view. . . 78

4.4 Molecular orbitals of N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine in the equatorial conformation. HOMO-2 top view (top, left) and side view (bottom, left). HOMO-20 top view (top, right) and side view (bottom, right). 79
4.5 Relaxed potential energy scan of the rotational barrier of R,R-N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine in the equatorial (circle) and axial (square) conformation around the C-N dihedral between the five and six membered rings. Continuous lines are spline fits to the data. 80
4.6 Temperature dependent proton NMR of R,R -N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine at 25 °C, 50°C, 65°C, and 80°C with labeled peaks. 82
4.7 Conformational changes between the four stereochemistries of meso N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine (2c). Red arrows indicate ring rotation. Blue arrows indicate pucker transitions. 83
5.1 Model RNA transesterification reaction. 85
5.2 Phosphate backbone of RNA with torsions labeled. 91
5.3 *Ab initio* (top frame) and CHARMM (bottom frame) torsional potential energy surface for the C-O-P-O dihedral of dimethyl phosphate (OO), non-bridging thiosubstituted dimethyl phosphate (SO), and non-bridging dithiosubstituted dimethyl phosphate (SS). 92
5.4 CHARMM27 standard and modified nucleotide residues 108
5.5 Complexes used for Mg^{2+} and OH^- Parameterization 109
5.6 Complexes used for S and Mg^{2+} Parameterization 110
5.7 Complexes used for Thio-substituted Phosphorane Parameterization . . 111

This dissertation contains work previously published and is reproduced with permission.

- Chapter 3: With kind permission from Springer Science Business Media: "Density Functional Study of the Influence of C5 Cytosine Substitution in Base Pairs with Guanine" Moser, A. Guza, B. N. Tretyakova, D. M. York, *Theor. Chem. Acc.* (2009) 122:179-188. DOI 10.1007/s00214-008-0497-5
  Licence Number 221038073512

- Chapter 5.1

  With kind permission from John Wiley & Sons, Inc. "CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes" Mayann E., Moser A., MacKerell Jr A. D., York D. M. *J. Comput. Chem.* (2007) 28: 495-507 DOI 10.1002/jcc.20474

Chapter 1

Modeling biological systems through computational chemistry

1.1 Computational chemistry

Computational chemistry is a sub-discipline of chemistry that develops and utilizes computer models of chemical systems. As computer power has increased over the last half decade, a variety of models have been created to investigate all types of chemical questions. Because of the relative expense of chemicals and analytic equipment compared to computer power, the inherent safety concerns of working with either carcinogenic, explosive, or toxic chemicals, and the absolute control within the computer, computational methods provide a valuable contribution to scientific research.

Any modeling, including computational chemistry, has a very particular relationship with experiment. Experiment often produces very specific information like NMR chemical shifts, absorptions or emissions spectrum, melting temperatures, etc. However, these experimental observables do not usually lend themselves to an unambiguous chemical interpretation (e.g., in terms of structure, mechanism, etc...) without the aid of models. Spectra need to be assigned, electron density needs to be fitted, linear free energy relationships need to be modeled, and so forth. Computational models can help to bring together the primary data provided by experiments and provide a complete chemical picture. At the same time, computational models rely on experimental data

for benchmarking, parameterization, and verification so that reliable predictions and results are obtained. In this way, modeling and experiment have a symbiotic relationship with computational chemistry using experimental results to motivate questions and provide benchmark results and computational chemistry providing prediction, elucidation to those questions, and motivation for future experimental work.

All models are a particular mix of approximations, cost, accuracy, precision, and applicability. Choosing the appropriate model that can reliably provide insight into a problem within the limits of the computational resources available is the first step in any computational chemistry project. This work presents several applications of computational methods to a variety of biochemical problems, specifically those related to nucleic acids. For each problem, models are chosen that balance cost and predictive capacity to provide the reliable results for the properties of interest.

1.2 Challenges in modeling biomolecules

Nucleic acids, proteins, lipids, and carbohydrates are the building blocks of life. From a computational chemistry prospective these are very challenging systems for a variety of reasons. First, as chemists our main unit for describing systems is the atom, making biochemical systems extremely large. A cell can be estimated to include around 10^{14} atoms, which is far beyond any atomistic chemistry model. Even more simplistic biochemical systems, for example RNP complex[1] or Group II Intron,[2] are at the very edge of most atomistic computational models. Compounding the size problem of most biochemical systems, is that research is rarely interested in them in a vacuum. The biochemical environment is very complex including counter ions, cofactors, other biomolecules, water, etc. For example, if the system of interest is a membrane protein, one would likely need to model the lipid bilayer and aqueous environment (i.e. water, ions, etc) as well as the protein. These environments create systems that are sensitive to their non-bonded interactions, requiring accurate representation of the electrostatic, dispersion, and repulsive forces and further increase the size of the system. The third challenge in modeling is conformational variation. Nucleic acids and proteins are biopolymers that, while having specific folded states, still have significant conformational freedom. When problems that require these conformations to be considered, significant computational resources

are required to sample this phase space, which may include millions of structures. Last, chemistry is most often interested in reactions, the breaking or forming of chemical bonds. To accurately represent this process, models must include some representation of the electronic structure. This increases the complexity of the model as well as the computational expense.

The primary motivation of research in the York Lab has been focused on understanding the structure, dynamics and mechanism of catalytic RNA, so called ribozymes. Understand of these biocatalysts can be leveraged into medical therapies,[3–18] as well as other biotechnologies;[19–23] specific example are ribozyme controlled allosteric switches[24–34] and biosensing devices.[35–40] This places understanding of nucleic acid structure, dynamics, and reactivity as a priority.

Nucleic acid systems present all the modeling complications described above. In particular, a ribozyme like spliceosome is a large, complex system comprising proteins and nucleic acids.[41–43] Smaller ribozymes (e.g. hepatitus delta virus,[44] hairpin,[45] hammerhead[46]) present significant conformational freedom and carry a negative charge for each nucleotide from the phosphate backbone, requiring long time scale simulation and extensive ion and solvation environments.[47] Because ribozyme systems are so complicated, no single modeling technique is sufficient to answer all the structural and mechanistic questions. Rather, a combination of a variety of computational methods are required.

What follows is a brief background of some of the methods used in this work and how they can be applied to biochemical problems. The intent here is to provide a concise summary of the key methods to provide the necessary background to understand how the methods fit together in a multiscale modeling framework to solve greater problems.

1.3 Electronic structure

The central equation of quantum chemistry is the time independent Schrödinger equation

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \tag{1.1}$$

where $\hat{H}$ is the Hamiltonian operator, $|\Psi\rangle$ is the wave function in Dirac notation, and E is the energy. The wave function completely describes a system and can provide any physical observable given the proper operator.[48] Equation 1.1 can be simplified

in a variety of ways including the Born-Oppenheimer approximation to separate the nuclear and electronic degrees of freedom, and disregarding relativistic corrections, but still leaves the complication of solving a many-electron wave function.

The wave function is often modeled as product of single particle, in this case electron, spin orbitals. In computational chemistry, these functions are called molecular orbitals and are composed of a spatial part that describes the spatial distribution of an electron, $\phi(r)$, and a spin part, $\chi(\omega)$, with spin state ω

$$\psi_i = \phi_i(r)\,\chi_i(\omega) \tag{1.2}$$

Further, $|\Psi\rangle$ must conform to the antisymmetry principle that states the wave function must change sign on the exchange of any two electrons, which can be accomplished by representing the wave function as a Slater determinant of single electron wave functions, ψ_i. In Dirac notation this is shown as

$$\begin{aligned} |\Psi_{ele}\rangle &= |\psi_i\psi_j\cdots\psi_k\rangle \\ &= \frac{1}{\sqrt{N!}}\sum_{n=1}^{N}(-1)^{p_n}P_n(\psi_i(1)\psi_j(2)\cdots\psi_k(N)) \end{aligned} \tag{1.3}$$

where $|\Psi_{ele}\rangle$ is the wave function of the electrons, N is the number of electrons, $\psi_i(j)$ is the i^{th} orthonormal spin orbital occupied by electron j, P_n is a permutation operator that interchanges the coordinates of electrons, n is an index of permutation, and p_n is the integer number of elementary coordinate exchanges in the ${P_n}^{th}$ permutation.

The Hamiltonian operator, when only considering the energy of the electrons ($\hat{H}_{ele}$), can be written (in atomic units) as

$$\hat{H}_{ele} = \sum_{i=1}^{N} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} - \sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{r_{ik}} \tag{1.4}$$

where i and j go over the total number of electrons, N, separated by distance r_{ij} and k goes over the total number of nuclei, M, with atomic charge Z_k. The three summations of Equation 1.4 are the electron kinetic energy operators, electron-electron potential energy operators (repulsive), and electron-nuclei potential energy operators (attractive). The electronic Hamiltonian is central in determining electronic structure, as $\langle\Psi_{ele}|\hat{H}_{ele}|\Psi_{ele}\rangle$ returns the electronic energy of the system and the minimization of this energy, under the proper normalization and antisymmetry constraints, determines the wave function $|\Psi_{ele}\rangle$.

1.3.1 Hartree-Fock and Semiempirical

Hartree-Fock (HF) is a common method to obtain a many-electron wave function.[49] This method, usually formulated in the matrix representation of Roothaan,[50] relies on assuming an average electron potential being felt by each individual electron and results in a self-consistent field procedure to solve.[51] Due to this mean field assumption, it neglects electron correlation, which can be problematic when dealing with more complicated biomolecular systems (e.g. base stacking or hydrogen bonding). Beyond the inherit approximations of HF, a representation of the molecular orbitals must be chosen, which is often a set of mathematical functions usually based on atomic orbitals. A complete set of basis functions is not possible (because it would require an infinite number of them), so some smaller basis set must be used, limiting the accuracy of the calculation. While the basis set can by systematically improved by adding more functions, this comes at a cost. HF formally scales as N^4, where N is the number of basis functions used in representing the wave function. This scaling limits its application to larger systems.

For application to large biomolecules it has been effective to implement semiempirical models that begin with the HF equations and replace some of the most expensive parts of the computations with parameters. This speeds up the calculation, but requires the parameters to be chosen with care such that the calculations still produce reliable predictions. Further, these semiempirical methods often employ a minimal (valance electron only) basis set to represent the molecular orbitals, leading to the minimal computational cost. General parameterizations have been done in multiple forms.[52–54] To apply semiempirical models to specific biological questions, it is can be useful to expand on them, as with expanded basis sets,[55] by tuning the parameters specifically for the particular problem,[56,57] or by decreasing their scaling.[58] These parameterizations are a mixture of chemical intuition, mathematical optimization, and careful choices on what benchmark data is used. So while the form of the semiempirical models can by systematically improved, all require some creative and thoughtful attention.

This work does not utilize HF theory or semiempirical models directly, so no more detail is required. Parameterization of new semiempirical models is a goal of Chapter 2, so a general idea of the model is useful. The calculations in this work utilize another electronic structure method that overcomes some of the hurdles of HF theory, and is

the focus of the next section.

1.3.2 Density functional theory

An alternative way to obtain information on the electronic structure is by density functional theory (DFT). This method utilizes the electron density, $\rho(\mathbf{r})$, as the central variable, where $\mathbf{r}$ are the $3N$ Cartesian coordinates of the electrons. Rather than the wave function, the energy is expressed as a functional (a function that takes a function as an argument and returns a scalar) of this density

$$E[\rho] = F[\rho] + \int \rho(\mathbf{r})\nu(\mathbf{r})d^3\mathbf{r} \tag{1.5}$$

where $E[\rho]$ is the energy functional, $F[\rho]$ is the kinetic energy and electron-electron interaction energy functional, and $\nu(\mathbf{r})$ is an external potential (usually the atomic nuclei). This method is based on two theorems of Hohenberg and Kohn (HK).[59] The first, known as the existence theorem, is that there is a unique mapping between the ground state electron density and the external potential (often the nuclear attraction and electron repulsion in a molecule). The result of this theorem is that there exists some functional of ρ that returns the exact ground state energy. Second, there exists a variational principle on the energy for any trial density, $\widetilde{\rho}$, relative to the energy of the true ground state electron density, ρ_0,

$$E[\rho_0] \leq E[\widetilde{\rho}] \tag{1.6}$$

given a non-zero density and a fixed number of electrons

$$\widetilde{\rho}\,(\mathbf{r}) \geq 0 \;\forall\; \mathbf{r} \tag{1.7}$$

$$\int \widetilde{\rho}\,(\mathbf{r})\, d^3\mathbf{r} = N \tag{1.8}$$

where N is the total number of electrons in the system. This leads to a constrained minimization procedure based on

$$\delta\left\{E[\rho] - \mu\left(\int \rho(\mathbf{r})d^3\mathbf{r} - N\right)\right\} = 0. \tag{1.9}$$

where the Lagrange multiplier enforcing the constraint is the electronic chemical potential.[60] Equation 1.9 can be recast into equations similar to HF theory by allowing

part of the energy to be represented by molecular orbitals leading to a self-consistent formulation where the density is optimized to produce the lowest system energy. This is known as the Kohn-Sham self-consistent field method (KS-DFT) and is the basis of the DFT calculations in this work. For more details on the development of DFT see References 60 and 61.

Unlike HF, DFT provides the possibility of an exact solution to the electron density and system energy (it does not *a priori* neglect electron correlation). The problem is that the form of the energy functional that achieves this is unknown. As stated above, $F[\rho]$ is a sum of the electron kinetic energy and electron-electron interaction,

$$F[\rho] = T[\rho] + J[\rho] + E_{QM}[\rho] \tag{1.10}$$

where $T[\rho]$ is the kinetic energy functional, $J[\rho]$ is the classical electrostatic energy functional, and $E_{QM}[\rho]$ is the electron interactions based on their quantum mechanical nature (e.g. exchange and correlation). Only the functional form of $J[\rho]$ is known, and nothing in the HK theorems provide insight into the functional form of $T[\rho]$ or $E_{QM}[\rho]$. Further, once an approximation for $F[\rho]$ is made and tested, there is rarely a systematic way to improve its reliability.

The development in DFT methods has revolved around approximations of $F[\rho]$ that best predict various physical properties. Most commonly for development, $F[\rho]$ is written as

$$\begin{aligned} F[\rho] &= T_s[\rho] + J[\rho] + (T[\rho] - T_s[\rho] + E_{QM}[\rho]) \\ &= T_s[\rho] + J[\rho] + E_{xc}[\rho] \end{aligned} \tag{1.11}$$

where $T_s[\rho]$ is the kinetic energy functional of the non-interacting electrons and $E_{xc}[\rho]$ is the exchange-correlation functional, which contains the correction to the kinetic energy based on the electron-electron interaction as well as all non-classical electron-electron interactions. Most work has led to increasingly more complicated forms of $E_{xc}[\rho]$ that have had great success in a variety of applications. For an excellent overview and comparison of various functionals see Reference 51. The work here mostly focuses on the three parameter hybrid B3LYP[62–64] and will be highlighted and detailed in multiple chapters.

1.3.3 Multilevel methods

While HF is limited by its neglect of electron correlation, various post-HF methods overcome these limitations. These methods include Møller-Plesset perturbation theory, coupled-cluster theory, configuration interaction, etc. All of these provide significant improvement in accuracy but at a substantial computational cost due to their greater scaling in the number of basis functions and intrinsic cost. In an effort to keep cost low while retaining accuracy, many multilevel model chemistries have been developed. Multilevel methods try to extrapolate the post-HF corrections and the infinite basis set limit by a series of calculations. For example, G2, developed by Pople *et al.*[65] was one of the first of these types of methods. The method includes

- HF/6-31G(d) geometry optimization
- HF/6-31G(d) frequencies for zero-point vibrational energy
- MP2(full)/6-31G(d) final geometry optimization
- MP4/6-311+G(d,p) and MP4/6-311G(d,p) to extrapolate the effect of diffuse basis functions
- MP4/6-311G(2df,p) to extrapolate the effect of higher angular momentum basis functions
- QCISD(T)/6-311G(d) to extrapolate more accurate correlation description
- MP2/6-311+G(3df,p) base energy, largest basis set

Each of these calculations are then combined in order to provide an estimate of the electronic structure properties in the infinite basis set and theoretical model limits. This leads to very accurate results, < 1 kcal/mol average error in atomization energy. At the same time this calculation is not as expensive as a full geometry optimization and energy evaluation at QCISD(T) at the largest basis set used. In this way, multilevel methods are able to approach chemical accuracy (1 kcal/mol), while having a reasonable cost for small molecules.

1.4 Molecular mechanics

As discussed above, biological systems can be very large and even the fastest, best scaling semiempirical models can only handle on the order of a few thousand atoms and this for a single or few energy evaluations.[66] For problems that require sampling many configurations of a large system, molecular mechanics force fields are often employed. These force fields forgo explicit description of the electronic structure and describe the energy of a system based on a summation of terms that represent the bonded and non-bonded atomic interactions. This approach is significantly less computationally expensive than even the cheapest electronic structure methods. A force field potential energy function may take the form of

$$\begin{aligned} U &= \sum_{bonds} k_b(r - r_{eq})^2 + \sum_{angles} k_a(\theta - \theta_{eq})^2 \\ &+ \sum_{dihedrals} k_\chi[1 + \cos(n\chi - \chi_{eq})]^2 \\ &+ \sum_{i,j<i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{r_{ij}} \end{aligned} \tag{1.12}$$

where the first three summations describe the bonded interactions and the final two summations describe the nonbonded interactions.

The bond energy terms sum over all the chemical bonds, angles, and dihedrals. The bond (two-body) and angle (three-body) terms are the same form as Hooke's Law, with the quadratic of the displacement about some equilibrium value (denoted eq) multiplied by a force constant (k). The dihedral (four-body) term is slightly different as it is periodic, giving the extra parameter n to provide the correct periodicity. The nonbonded term has two parts. The first describes repulsion and dispersion, which are representations of the electronic structure phenomena correlation and exchange repulsion, respectively. The leading attractive dispersion term goes as r^{-6} and the repulsive exchange goes as r^{-12}. This empirical form for these two interactions is known as the Lennard-Jones (LJ) potential, where $-\epsilon_{ij}$ is well depth and σ_{ij} is the contact distance of zero energy. The second part of the nonbonded term is Coulomb's Law, representing the electrostatic interaction of all the atoms. For this term some atomic charge (q_i) must be defined for each atom. Note that the number of bonded interactions is much

fewer than the number of nonbonded interactions for large systems as every atom may interact with every other atom in the nonbonded terms, while the bonded terms only see a local region of the system. This results in much more computational effort evaluating these longer range, nonbonded terms than the bonded terms.

Equation 1.12 in theory requires a significant number of parameters: k_b, k_a, k_χ, ϵ_{ij}, σ_{ij}, q_i for each bond, angle, dihedral, atom, and atom pair. This can be simplified significantly as experience and experimental data have shown that many chemical situations are highly transferable when it comes to values of the force constants and equilibrium values.[51] For example, the k_b and r_{eq} for the C-C bond of ethane is very similar to the two C-C bonds in propane. We can apply this to the LJ term as well by making ϵ_{ij} and σ_{ij} a function of single atom parameters. So in this way we create transferable atom types to describe similar situations and dramatically decrease the number of parameters and time spent obtaining them. Similar to semiempirical parameterization, what you parameterize your force field to has a significant effect on the outcome and just like density functional development this has led to a large variety of force fields[51] each with its own ranges of reliability and use.

Molecular mechanics force fields, due to their computational efficiency, can be applied to very large systems and used to create ensembles of structures through time propagation of Newton's equations (molecular dynamics) or by Monte Carlo sampling.[67] One of the drawbacks of this model is that while these force fields have had great success describing the structural aspects of many systems, most do not allow for bond formation or cleavage.

1.5 Solvation

The importance of water in biological systems cannot be overestimated. All life on earth is in some way intrinsically linked to its aqueous environment. Therefore when biomolecular questions arise, the solvent must be considered carefully. Not only must water be well described, but most systems have significantly more water than solute. A 1 μM solution of DNA has 55 million water molecules for every 1 DNA molecule. Computationally, this is too many waters to describe in an explicit way, so various schemes to reduce the number, like periodic boundary conditions (PBC),[67] and increase

the evaluation speed of these interactions, like fast multipole methods and Particle-Mesh Ewald[68,69] are used. Even with these developments, the total number of water scales with the size (and sometimes complexity) of the solute and are still too numerous to be described in detail for many biomolecules.

From a computational point of view solvation has been mainly modeled in two ways: explicitly and implicitly. Explicit solvation involves creating atomistic representations of the water molecule, but as noted above due to the number of waters needed, this model must be computationally expensive. A variety of molecular mechanic water models have been developed for this purpose.[70–73] When a system is highly sensitive to solvent structure (e.g. hydrogen bonding analysis), explicit solvent provides a computationally tractable way of modeling these interactions.

Implicit solvation models attempt to represent the solvent effect on the solute in some average way. Many of these models' goal is to predict the free energy of solvation (ΔG_{solv}), which is the free energy of moving a solute from gas-phase to condensed phase and important in determining partition coefficients. Often the solvent interactions are separated into electrostatic and nonelectrostatic contributions

$$\Delta G_{solv} = \Delta G_{ele} + \Delta G_{nonele} \tag{1.13}$$

where ΔG_{ele} is the solvation free energy from the electrostatic interaction of the solute with solvent and solvent with itself and $\Delta G_{nonelec}$ is all other contributions.

The electrostatic contribution can obtained by representing the solute as some charge distribution and embedding it within some continuum dielectric, see Figure 1.1. This requires the definition of some boundary that separates the solute and solvent, often based on some atomic radii that are then combined to form a solute surface. ΔG_{nonele} can be further broken down as

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{dis} + \Delta G_{rep} \tag{1.14}$$

where ΔG_{cav}, ΔG_{dis}, and ΔG_{rep} are the cavitation, dispersion, and repulsion contributions, respectively. Cavitation is the free energy penalty to moving the water out of the solute cavity. Dispersion and repulsion are synonymous to that described in the molecular mechanics section, in this case between the solute and solvent.

A wide range of solvent models are based on continuum dielectric methods have been developed;[74–77] most common are finite difference Poisson-Boltzmann,[78] multipole

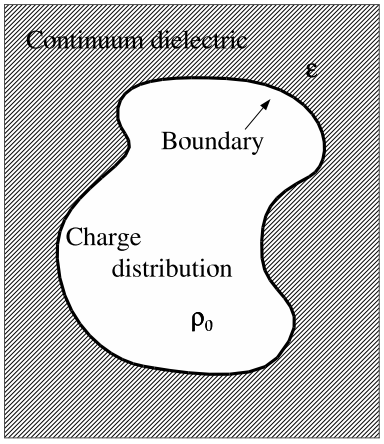

Figure 1.1: Charge distribution (ρ_0) in a continuum dielectric (ϵ).

expansion (MPE),[79] polarizable continuum model (PCM),[80,81] conductor-like screening model (COSMO),[82–84] and the SMx models.[85–87] Each implicit solvation model goes about describing this solvation energy slightly differently, but all are highly parameterized models using experimentally known solvation free energies.

Both explicit and implicit models have seen significant application to biochemical systems. Explicit models are often used at a greater computational expense, but provide explicit solvent structure about the region of interest. Implicit models do well at describing the bulk solvent effect for less computational cost and are highly parameterized to reproduce solvation free energies, but do not provide specific solvent structure.

1.6 Multiscale modeling

While quantum methods can accurately describe bond breaking and forming, they are too computationally expensive to apply to most biochemical systems of interest that are many thousand atoms. On the other hand, molecular mechanics methods are able to accommodate these larger systems, but do not describe the electronic structure required to model reactivity. Further, accurate long range solvation is necessary to accommodate the highly charged nucleic acid systems. The strategy undertaken in the York Lab is to both advance these individual computational components, but also design and combine

various techniques into multiscale models.

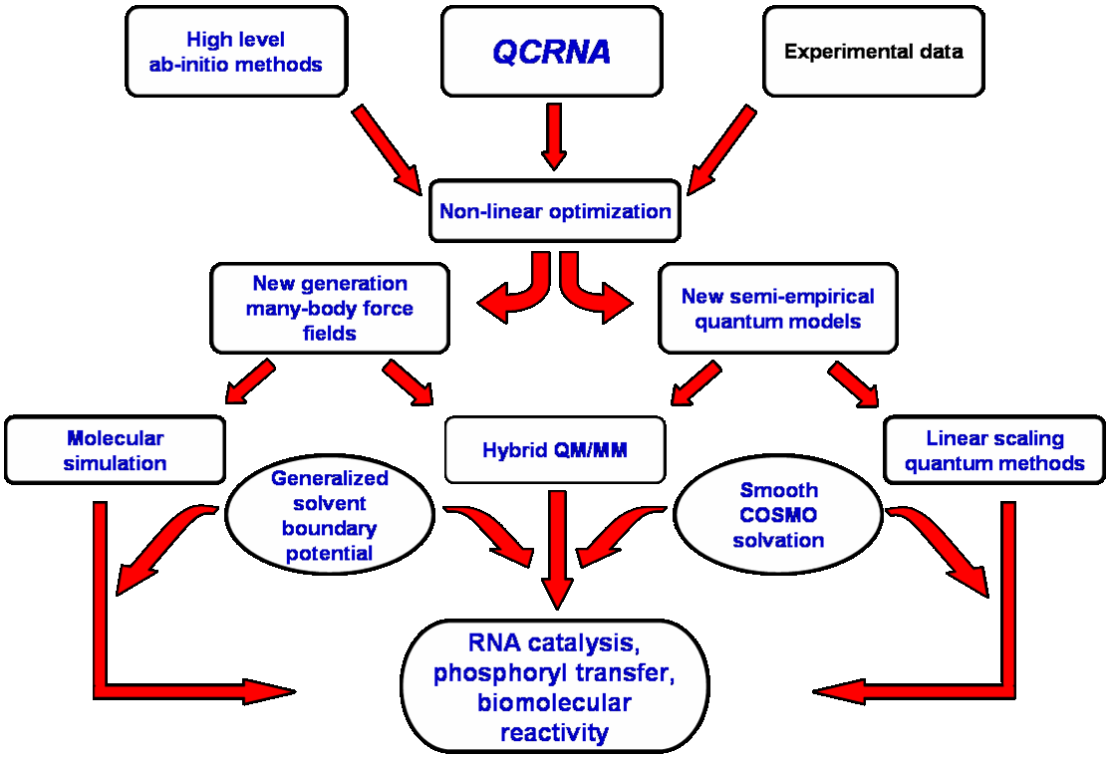

Figure 1.2: Schematic flow of various computational methods and experimental data used in the development of multiscale models.

Figure 1.2 is a schematic of how various computational methods can fit together to develop multiscale models for RNA catalysis, phosphoryl transfer, and biomolecular reactivity. At the top are high level *ab initio* methods (e.g. multilevel methods) and experimental data, that can be used as the "gold" standard for benchmarking less expensive methods. At times neither experimental data or high level *ab initio* calculations are available or feasible. This is why the top level also includes QCRNA, which is a large database of density functional calculations of relevant structures, potential energy surfaces, and small molecule reactions.[88] While this database is not as generally reliable as the other two options, by looking at trends and differences over the large data set useful information can be obtained. In combination these three sources provide

a thorough benchmark of data to parameterize, through non-linear optimization, next generation force fields and semiempirical quantum models. Force field parameterization is essential for accurate sampling of large biomolecules. New semiempirical quantum models are a computationally inexpensive, yet reliable once parameterized, method to describe chemical reactions.

While force fields can be used directly in molecular dynamics simulations and semiempirical models can be developed with linear scaling and applied to larger systems, they can be combined in hybrid quantum mechanical/molecular mechanical (QM/MM) simulation. QM/MM applies quantum mechanics to the reactive site and the less expensive molecular mechanics to the rest of the system. In this way, the more expensive computational model is applied just to the atoms that especially require it. To help accommodate a more complete solvation and/or biomolecular environment, generalized solvent boundary potentials can be used to envelope QM/MM system. Further, implicit solvation models can be used in conjunction with explicit solvent models to provide both local solvent structure and bulk solvent effect.[89,90]

Development of these combined multiscale models is a complicated process because not only must each individual method be verified but the boundaries between each model must be careful considered. Once these methods are fully developed, they are able to model significantly larger systems than a quantum mechanical description, while still describing biomolecular reactivity.

1.7 Preview

The following chapters contain application of computational chemistry techniques to biomolecular systems. Chapter 2 presents a benchmark investigation into the modeling of gas phase protonation and deprotonation events using multilevel methods and density functional methods. Then these methods are employed to predict the proton affinity and gas-phase basicity of amino acids both side chain and backbone, nucleic acids in both the keto and enol tautomers, phosphates with and without thio-substitution, and RNA sugar. These results provide valuable and comprehensive data for the parameterization and optimization of future semiempirical models designed for hybrid QM/MM investigations of biocatalysis and provide insight into the intrinsic reactivity of these

sites, which may guide further biochemical research.

In Chapter 3, an investigation of how cytosine methylation at C5 can affect the reactivity of a base paired guanine to carcinogen attack is presented. Some current experiments in this area have employed the use of analogues to investigate the role of the methyl group in the enhancement of carcinogen reactivity. By replacing the methyl group with various other substitutions, it is hoped that the role of the methyl can be clarified. The purpose of the computational work is to help guide the experimental results and clarify various steps within the mechanism.

Chapter 4 is a brief look on exocyclic lesions formed during carcinogen attack at adenine. This work is a mixture of experiment performed in the Tretyakova Lab and computational work by the York Lab. Computational chemistry is employed in various ways to elucidate NMR results, understand structure features, and rationalize product distributions.

While the first three chapters focus on quantum mechanical investigations of various biochemical questions, these results also provide a benchmark for the parameterization of computationally faster models that can be applied to larger systems. Chapter 5 provides a glimpse into how various quantum computational results can be harnessed for parameterization of molecular mechanics models. This work includes parameterization of transition state mimics for phosphate hydrolysis and investigation of reference and computational data for phosphate solvation.

Chapter 2

Proton affinity and gas-phase basicity of biocatalytic molecules

2.1 Introduction

The charge state of proteins and nucleic acids play an important role in both structure and reactivity. A key mechanism of controlling charge state and influencing acid-base catalysis is through protonation/deprotonation events of ionizable residues.[91] There has been considerable effort devoted to the prediction of pK_a values with quantum chemistry and implicit/explicit solvation models.[92–115] Much of the quantitative work has focused on the prediction of pK_a shifts of ionizable residues with respect to a closely related reference state (for which reliable absolute pK_a values have been determined) rather than on the prediction of the absolute pK_a values themselves.

When carefully tested, this type of indirect approach often leads to cancellation of systematic errors that ultimately results in more quantitative accuracy. Examples of commonly used reference state for calculation of pK_a shifts include a closely related molecule or residue, or else the same molecule or residue in a different environment. The absolute pK_a values can be recovered from the calculated pK_a shift and the experimentally known absolute pK_a value of the reference state. Nonetheless, this area remains a challenge due to the small differences in free energy that give rise to pK_a shifts (a pK_a unit corresponds to 1.364 kcal/mol of energy at 298.15 K).

Reliable prediction of pK_a shifts using known experimental data may not always be

possible, since the determination of the experimental pK_{a} value of an appropriate reference state might not be available, e.g., metaphosphates and phosphoranes that form reactive intermediates in phosphoryl transfer reactions. For these important intermediates, experiment can only provide a rough estimate of the pK_{a} values. Consequently, methods that improve the prediction of absolute pK_{a} values are also of importance, and ultimately, may lead to the determination of pK_{a} shifts with increased reliability.

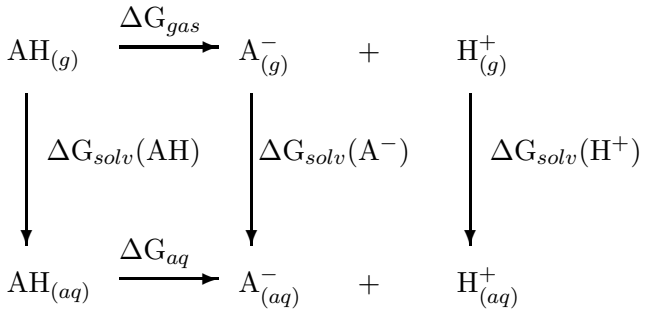

Figure 2.1: A commonly employed thermodynamic cycle for calculating pK_{a} values.

The accurate prediction of absolute pK_{a} values is often made using a thermodynamic cycle such as the one shown in Figure 2.1. A key quantity in the cycle is the gas-phase basicity (GPB) that involves deprotonation of the species of interest in the gas-phase (ΔG_{gas} in Figure 2.1). Related to the gas-phase basicity is the proton affinity (PA), which is the enthalpic component of the same process.[116] Given the active site of most enzymes or the center of large biomolecules is an environment that is neither that of bulk solution nor gas-phase, benchmarking these two end point environments is key to making pK_{a} predictions.

The purpose of the present work is to present consistent, benchmark quantum calculations for proton affinities and gas-phase basicities of residues involved in biocatalysis. Compounds were considered that represent titratable amino acid backbone and side-chain residues, nucleic acid bases, sugar and phosphate moieties as well as biological and chemically modified phosphates and phosphoranes important in phosphoryl transfer reactions and RNA catalysis. The accuracy of a series of high-level quantum model

chemistries was assessed against known experimental proton affinity and gas-phase basicity values. This comparison was then used to derive a set of empirical bond enthalpy and free energy corrections and a linear regression correction that improves the accuracy and predictive capability of the methods.

The results of the present work are instrumental for the development of new semiempirical quantum models for combined QM/MM calculations of biocatalysis,[117] understanding the relative protonations of biomolecules and their implication to reactivity and binding,[118–120] comparison to current experimental estimates,[121] and the prediction of pK_{a} and solvation free energy values of biological residues for which experimental data is not available.[103]

2.2 Methods

2.2.1 Electronic structure calculations.

All electronic structure and thermochemical analysis was performed using the Gaussian03 suite of programs.[122] Three "multi-level" methods were studied: CBS-QB3,[123,124] G3B3,[125] and G3MP2B3.[125] CBS-QB3 (abbreviated CBS in this article) is a multi-level model chemistry that combines the results of several electronic structure calculations and empirical terms to predict molecular energies to around 1 kcal/mol accuracy.[126] The required electronic structure calculations are outlined below (see Ref. 123 for details):

CBS:

- B3LYP/6-311G(2d,d,p) geometry optimization and frequencies
- MP2/6-311+G(3df,2df,2p) energy and CBS extrapolation
- MP4(SDQ)/6-31+G(d(f),p) energy
- CCSD(T)/6-31+G† energy

The G3B3 and the related G3MP2B3 (abbreviated G3MP) methods are both modifications of the Gaussian-3 multi-level theory for the calculation of molecular energies.[127] Like CBS-QB3, G3B3 and G3MP use density functional theory with the B3LYP functional for geometries and frequencies and combine the results of several electronic structure calculations and empirical terms to predict molecular energies to around 1 kcal/mol

accuracy. The required electronic structure calculations for the G3B3 method are outlined below (see Ref. 125 for details):

G3B3:

- B3LYP/6-31G(d) geometry optimization and frequencies
- MP2/G3Large energy
- MP4/6-31G(d) energy
- MP4/6-31+G(d) energy
- MP4/6-31G(2df,p) energy
- QCISD(T)/6-31G(d) energy

The G3MP method eliminates all of the MP4 calculations above, trading some accuracy for speed.

All of the multi-level methods studied here formally scale as $O(N^7)$ due to the CCSD(T) step in CBS, the QCISD(T) step of the G3 methods, and the MP4 steps in G3B3. (For the molecules in the present study, the large basis set MP4 step was usually the computational bottleneck with the G3B3 method). For the systems studied in this paper the scaling of the multi-level methods was not prohibitive, but for larger systems of biological interest the $O(N^7)$ will eventually dominate and make the calculation unfeasible. Therefore several model chemistries based solely on hybrid density functionals were investigated that have more favorable scaling properties, specifically PBE0/6-311++G(3df,2p) (PBE0), B1B95/6-311++G(3df,2p) (B1B95), and B3LYP/6-311++G(3df,2p) (B3LYP). Additionally, the B3LYP/6-311++G(3df,2p)//B3LYP/6-31++G(d,p) model chemistry (designated QCRNA) that has been extensively applied to model biological phosphorous compounds[88,97,128–130] was also included.

B3LYP is a three parameter hybrid functional[63] using the B88 exchange functional of Becke[62] and the Lee, Yang, and Parr correlation functional.[64] PBE0 is a zero parameter hybrid functional[131] using the Perdew, Burke, Ernzerhof exchange and correlation functional.[132] B1B95 is a one parameter hybrid functional[133] using the B88 exchange functional of Becke[62] and B95 correlation functional.[133]

For some molecules, especially those containing sulfur, the default implementation of the B95 functional in Gaussian03 exhibited SCF convergence problems. In those cases, B1B95 calculations were done with a version of Gaussian[122] modified to avoid

the underlying numerical instability in the evaluation of the B95 functional.[134] The more stable implementation of B95 produces energies that differ by $\approx 10^{-5}$ Hartrees (≈ 0.006 kcal/mol) for the phosphorous and sulfur containing molecules in this work.

In an effort to give the hybrid density functional methods the best possible accuracy for benchmark purposes and avoid problems in the frequency calculations[135] all calculations (except the QCRNA and multilevel theories) were run with the so-called "ultrafine" numerical integration grid (a pruned grid based on 99 radial shells and 590 angular points per shell) and tight convergence criteria for the geometry optimizations. The use of large basis sets, ultrafine numerical integration meshes, and tight convergence criteria serve to make these calculations benchmark quality. The QCRNA model is significantly less expensive than the related B3LYP model since it avoids the geometry optimization and frequency calculation steps with the large 6-311++G(3df,2p) basis, and uses the default integration grid (a pruned grid based on 75 radial shells and 302 angular points per shell) and geometry convergence criteria. As demonstrated below, QCRNA gives results in very close agreement to that of B3LYP.

2.2.2 Calculation of proton affinities and gas-phase basicities.

The proton affinity (PA) and gas-phase basicity (GPB) of a species A^- are related to the gas-phase reaction:

$$A^-_{(g)} + H^+_{(g)} \rightarrow AH_{(g)} \tag{2.1}$$

The proton affinity of A^- is defined as the negative of the enthalpy change (ΔH) of the process in Eq. 2.1, and the gas-phase basicity of A^- is defined as the negative of the corresponding Gibbs free energy change (ΔG).[116] Here, the A^- is the conjugate base associated with the neutral acid AH.

The required thermodynamic properties were obtained from the electronic structure calculations using standard statistical mechanical expressions for separable vibrational, rotational, and translational contributions within the harmonic oscillator, rigid rotor, ideal gas/particle-in-a-box models in the canonical ensemble.[136] The standard state in the gas-phase was for a mole of particles at 298.15 K and 1 atm pressure.

In the case that a molecule has more than one conformational state accessible at a given temperature, these conformations should be Boltzmann averaged to obtain the

most accurate PA and GPB. In this work the free energies were used to determine the Boltzmann average and only conformations that changed the PA or GPB by more than 0.1 kcal/mol were considered. This occurred for propanol and propanethiol where the trans conformation is higher in energy than the gauche conformation but by less than 0.5 kcal/mol in free energy.

Explicit inclusion of zero-point energy corrections is important for reliable thermodynamic results. Although other studies have made the assumption that the difference in zero point energy between the neutral acid AH and anion A^- is generally small,[99] for the systems studied in the present work, this is not the case. An O–H bond stretch vibration typically falls in the range of 2500–3600 cm^{-1}, and corresponds to a zero-point energy difference of around 7–10 kcal/mol (this value is consistent with the zero-point energy range in the present work, e.g., 7.22–10.3 kcal/mol for the B3LYP method). Neglect of the zero-point energy, if applied to the calculation of pK_a values via the thermodynamic cycle in Figure 2.1 (or similar cycles), would lead to errors in absolute pK_a values of 5–7 pK_a units, and errors in pK_a shifts of up to 2 pK_a units. This would significantly limit the overall reliability and predictive capability in applications to biological systems, and hence it is recommended that zero-point energies be included explicitly in such calculations.

The enthalpy of the proton was calculated from the ideal gas expression,

$$H(\mathrm{H}^+) \;=\; U + PV = \frac{5}{2}RT \tag{2.2}$$

where U is the internal energy, P and V are the pressure and volume, respectively, R is the universal gas constant, and T is the absolute temperature. The entropy of the proton was calculated from the Sackur-Tetrode equation,[137]

$$S(\mathrm{H}^+) \;=\; R\ln\left(\frac{e^{\frac{5}{2}}k_BT}{p\Lambda^3}\right) \tag{2.3}$$

where k_B is the Boltzmann constant, p is the pressure, and Λ is the thermal De Broglie wavelength [$\Lambda = (h^2/2\pi m k_B T)^{1/2}$, where h is Planck's constant and m is the mass of the proton]. Under the standard state conditions, the values of the enthalpy and entropy of the gas-phase proton are $H(\mathrm{H}^+) = 1.48$ kcal/mol and $S(\mathrm{H}^+) = 26.02$ cal/mol K, respectively, and lead to a gas-phase Gibbs free energy value of $G(\mathrm{H}^+) = H(\mathrm{H}^+) - TS(\mathrm{H}^+) = -6.28$ kcal/mol.

It is sometimes the case that a molecule and/or its conjugate base has more than one indistinguishable microscopic protonation state. All of the GPB values in this paper are *microscopic* gas-phase basicities, since that is what naturally comes out of electronic structure calculations of a single protonation state. The conversion between microscopic and macroscopic GPB values was performed as follows.

The equilibrium constant, K^M, (where the M indicates macroscopic) for the reverse process to that of Eq. 2.1, assuming unit activity coefficients, is given by:

$$K^M = \frac{[A^-]_M[H^+]}{[AH]_M} \tag{2.4}$$

If A^- has m indistinguishable microscopic protonation states and AH has n indistinguishable microscopic protonation states, then:

$$K^M = \frac{m[A^-]_\mu[H^+]}{n[AH]_\mu} = K^\mu \left(\frac{m}{n}\right) \tag{2.5}$$

where μ indicates microscopic quantities and $K^\mu = [A^-]_\mu[H^+]/[AH]_\mu$. The free energy change (ΔG) for a process is related to the equilibrium constant by:

$$\Delta G = -RT \ln K \tag{2.6}$$

where R is the ideal gas constant and T is the temperature of interest. Substitution of Eq. 2.6 into Eq. 2.5 yields the following equation for interconversion of microscopic and macroscopic free energy changes:

$$\Delta G^\mu = \Delta G^M + RT \ln \left(\frac{m}{n}\right) \tag{2.7}$$

For example, H_3PO_4 has 4 indistinguishable microscopic protonation states (distribution of 3 protons among 4 sites gives ${}_4C_3 = 4$, where ${}_nC_k = n!/(n-k)!k!$ is a binomial coefficient) and its conjugate base, $H_2PO_4^-$, has 6 indistinguishable microscopic protonation states (${}_4C_2 = 6$). The experimental (macroscopic) GPB of H_3PO_4 is 323.0 kcal/mol.[121] Application of Eq. 2.7 yields a microscopic GPB of 323.2 kcal/mol at 298.15 K.

2.3 Results and Discussion

Unlike gas-phase or solution phase, a biological system (e.g. protein active site, within a DNA helix, along a phosphate backbone, etc) will likely have some perturbation in both

conformational freedom and electronic environment from either of these two states based on its relative position to rest of the system. To provide the most system independent values for the PA and GPB, we often resort to model compounds when conformational flexibility of an entire residue might greatly affect the calculations. For example, side chain conformation in methionine can tune the PA and GPB by ∼10 kcal/mol from the value for glycinium,[138] but it is unlikely that this same conformation is available in most biological systems. Here we focus on the intrinsic affinity of each site toward protonation or deprotonation so that direct comparison can be made both within in moiety type (i.e. amino acid, phosphate, etc) but also between them.

2.3.1 Experimental Comparison

Tables 2.1 and 2.2 summarize the PA and GPB errors (calculated - experimental value) for O-H, S-H, and N-H bond containing molecules. All experimental values are taken from the NIST online database[121] unless where otherwise noted. Included are the maximum error (MAXE), root mean squared error (RMSE), mean unsigned error (MUE), and mean signed error (MSE). As expected the multi-level methods preform better than the density functional methods for both PA and GPB, with PBE0 and B1B95 generally overestimating and B3LYP and QCRNA underestimating the values. As QCRNA consists only of B3LYP calculations, it very closely resembles the B3LYP results except for propanioc acid GPB where they differ by 1.8 kcal/mol. The largest outliers for all the multi-level models are dimethylphosphate and phosphoric acid. For PBE0 and B1B95, water and 4-methyl-imidazolium have the largest deviation, while B3LYP and QCRNA have maximum error on p-nitrophenol. For all models the PA and GPB generally have the same trends and error metrics.

In previous work for the PA and GPB of O-H containing compounds,,[139,140] it was found that for the DFT methods the |MSE| ≈ MUE, indicating a systematic error.[139] This observation motivated a simple additive correction for the bond enthalpy and entropy that brought the error metrics of the density function models closer to those of the multi-level models for both the PA and GPB, This PA correction (ΔH^C) is applied as

$$\Delta H'_X = \Delta H_X + \Delta H^C_X \tag{2.8}$$

where X is either O-H, S-H, or N-H bonds. This correction is similarly added for the

Table 2.1: Proton affinity error analysis. All quantities are in kcal/mol. Experimental values are from ref. 121 with error estimates given in parenthesis. Error is given as quantum model chemistry minus experimental value (error = calculated - experiment). Error metrics of maximum error (MAXE), root-mean-squared error (RMSE), mean unsigned error (MUE), and mean signed error (MSE) are given below.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA	Experiment
water	1.7	1.2	1.3	3.2	3.3	0.1	0.1	390.3 (0.2)
hydronium	-1.0	-0.3	-0.4	0.6	0.5	-1.1	-1.1	165.0 (1.0)
methanol	1.1	2.2	2.2	-0.5	-0.1	-2.3	-2.3	381.5 (1.0)
ethanol	0.7	1.5	1.7	-0.3	0.3	-2.2	-2.2	378.2 (0.8)
propanol	1.1	2.0	2.2	0.4	0.9	-1.2	-1.2	376.0 (1.1)
2-propanol	0.8	1.4	1.7	0.7	1.2	-1.4	-1.3	375.7 (0.8)
DMPH[b]	-2.4	-1.8	-1.3	-0.6	-0.7	-1.6	-1.5	331.6 (4.1)
phosphoric acid	-2.6	-2.2	-1.8	-0.5	-0.6	-2.4	-2.3	330.5 (5.0)
formic acid	-0.4	0.4	0.9	-0.2	0.1	-2.0	-2.0	344.0 (1.6)
acetic acid	0.2	1.0	1.4	0.4	0.7	-1.4	-0.8	347.2 (1.1)
propanoic acid	-1.2	-0.4	-0.0	0.4	0.5	-1.3	-1.9	347.4 (1.8)
glycine	1.5	2.3	2.7	2.7	2.9	1.2	1.2	340.3 (1.1)
proline	-0.2	0.7	0.9	1.8	1.8	0.1	0.1	340.3 (3.1)
phenol	-0.8	-0.6	-0.6	-1.4	-0.7	-2.4	-2.5	350.1 (1.1)
o-chlorophenol	0.7	1.0	1.0	0.4	1.0	-1.0	-1.0	343.4 (2.3)
m-chlorophenol	-0.5	-0.2	-0.2	-1.2	-0.6	-2.4	-2.3	342.1 (3.1)
p-chlorophenol	-0.5	-0.2	-0.2	-1.0	-0.5	-2.2	-2.3	343.4 (1.6)
p-methylphenol	-0.3	0.0	0.0	-0.7	-0.1	-1.8	-1.8	350.7 (1.3)
p-nitrophenol	-0.2	0.2	0.7	-2.3	-1.7	-4.1	-4.1	327.8 (2.1)
hydrogen sulfide	-0.5	-0.1	0.2	0.5	0.8	-0.3	-0.4	351.3 (0.8)
methanethiol	-0.5	0.1	0.4	0.9	1.4	0.0	-0.0	357.3 (1.2)
ethanethiol	-1.2	-0.5	-0.2	0.3	1.0	-0.4	-0.4	355.4 (1.5)
propanethiol	-1.0	-0.4	-0.1	0.6	1.2	0.2	0.2	354.2 (2.2)
2-propanethiol	-0.5	0.1	0.4	1.4	2.0	0.6	0.6	353.4 (2.2)
ammonium	0.2	0.5	0.3	0.6	0.2	-1.1	-1.1	204.0 (0.5)
methylammonium	0.1	0.6	0.4	0.5	0.2	-0.6	-0.6	214.9 (0.5)
ethylammonium	0.1	0.6	0.5	0.9	0.5	-0.1	-0.1	218.0 (0.5)
propanammonium	-0.1	0.4	0.2	0.9	0.7	-0.0	-0.1	219.4 (0.5)
2-propanammonium	-0.6	-0.1	-0.2	1.0	0.7	-0.1	-0.1	220.8 (0.5)
1H-3H-imidazolium	-0.0	0.7	0.4	1.9	1.8	1.3	1.3	225.3 (0.5)
4-methyl-imidazolium	1.3	1.9	1.7	3.6	3.5	3.0	3.0	227.7 (2.0)
pyridinium	-0.4	0.3	0.0	2.0	1.8	1.8	1.9	222.0 (5.0)
pyrrolidinium	0.1	0.8	0.7	1.0	0.8	0.6	0.6	226.6 (0.5)
glycinium	-0.0	0.4	0.3	1.7	1.3	0.1	0.1	211.9 (0.5)
MAXE	-2.6	2.3	2.7	3.6	3.5	-4.1	-4.1	
RMSE	1.0	1.1	1.1	1.4	1.4	1.6	1.6	
MUE	0.7	0.8	0.8	1.1	1.1	1.2	1.2	
MSE	-0.2	0.4	0.5	0.6	0.8	-0.7	-0.7	

[a] "Molecule" refers to AH in Eq. 2.1. [b] hydrogen dimethyl phosphate.

Table 2.2: Gas-phase basicity error analysis. All quantities are in kcal/mol. Experimental values are from ref. 121 with error estimates given in parenthesis. Error is given as quantum model chemistry minus experimental value (error = calculated - experiment). Error metrics of maximum error (MAXE), root-mean-squared error (RMSE), mean unsigned error (MUE), and mean signed error (MSE) are given below.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA	Experiment
water	1.7	1.2	1.3	3.2	3.3	0.1	0.1	383.7 (0.2)
hydronium	-1.1	-0.4	-0.5	0.5	0.3	-1.2	-1.2	157.7 (0.1)
methanol	1.3	2.4	2.3	-0.3	0.1	-2.0	-2.1	374.8 (0.7)
ethanol	0.4	1.3	1.4	-0.5	-0.0	-2.5	-2.4	371.3 (1.0)
propanol	0.7	1.6	1.8	-0.1	0.4	-1.8	-1.7	369.4 (1.1)
2-propanol	0.3	1.0	1.3	0.3	0.8	-1.8	-1.7	368.8 (1.0)
DMPH[b]	-1.7	-1.2	-0.7	0.2	0.3	-0.8	-0.9	324.6 (4.0)
phosphoric acid	-2.5	-2.2	-1.8	-0.3	-0.4	-2.2	-2.2	323.2 (4.9)
formic acid	-1.3	-0.6	-0.1	-1.1	-0.8	-2.9	-2.9	337.9 (1.2)
acetic acid	-2.0	-1.2	-0.8	-0.1	0.1	-1.9	-3.1	341.4 (1.2)
propanoic acid	-0.3	0.6	1.0	-1.2	-0.5	-2.8	-1.0	340.4 (1.4)
glycine	-0.4	0.5	0.8	0.7	0.9	-0.9	-0.8	335.1 (1.4)
proline	-0.9	-0.0	0.2	1.0	1.0	-0.8	-0.8	333.4 (3.0)
phenol	-0.6	-0.5	-0.5	-1.3	-0.7	-2.3	-2.4	342.9 (1.4)
o-chlorophenol	-0.7	-0.4	-0.4	-1.1	-0.5	-2.4	-2.4	337.1 (2.0)
m-chlorophenol	-1.2	-0.9	-0.9	-1.8	-1.2	-3.0	-2.9	335.2 (1.4)
p-chlorophenol	-0.9	-0.5	-0.5	-1.3	-0.7	-2.4	-2.6	336.5 (1.4)
p-methylphenol	-0.6	-0.4	-0.4	-1.0	-0.6	-2.0	-2.2	343.8 (1.2)
p-nitrophenol	-0.1	0.2	0.7	-2.3	-1.8	-4.1	-4.2	320.9 (2.0)
hydrogen sulfide	-0.5	0.0	0.2	0.5	0.8	-0.3	-0.4	344.9 (2.0)
methanethiol	-0.2	0.2	0.7	1.2	1.7	0.4	0.3	350.6 (2.0)
ethanethiol	-1.5	-0.0	-0.6	-0.1	0.7	-0.9	-0.8	348.9 (2.0)
propanethiol	-1.0	0.4	-0.2	0.5	1.0	-0.2	-0.2	347.9 (2.0)
2-propanethiol	-1.2	-0.9	-0.3	0.6	1.2	-0.1	-0.1	347.1 (2.0)
ammonium	0.3	-0.5	0.3	0.7	0.3	-1.0	-1.0	195.7 (0.5)
methylammonium	0.2	-0.6	0.6	0.5	0.3	-0.6	-0.6	206.6 (0.5)
ethylammonium	0.6	0.6	1.0	1.4	1.0	0.3	0.3	210.0 (5.0)
propanammonium	0.5	0.7	0.9	1.5	1.3	0.5	0.6	211.3 (0.5)
2-propanammonium	0.2	1.1	0.6	1.8	1.5	0.8	0.7	212.5 (0.5)
1H-3H-imidazolium	-0.5	1.0	-0.1	1.4	1.2	0.8	0.8	217.7 (0.5)
4-methyl-imidazolium	1.1	0.7	1.5	3.5	3.4	2.9	2.9	220.1 (2.0)
pyridinium	-0.8	0.1	-0.3	1.7	1.5	1.5	1.5	214.7 (0.5)
pyrrolidinium	0.4	1.7	1.2	1.3	1.1	0.9	1.0	218.8 (0.5)
glycinium	0.4	-0.0	0.7	1.7	1.5	0.3	0.5	203.7 (0.5)
MAXE	-2.5	2.4	2.3	3.5	3.4	-4.1	-4.2	
RMSE	1.0	1.0	0.9	1.3	1.2	1.8	1.8	
MUE	0.8	0.8	0.8	1.1	1.0	1.5	1.5	
MSE	-0.4	0.2	0.3	0.3	0.5	-0.9	-0.9	

[a] "Molecule" refers to AH in Eq. 2.1. [b] hydrogen dimethyl phosphate.

Table 2.3: Proton affinity error analysis for each model broken down by bond type and fitting method. Raw indicates no correction has been made. BEC and LRM are bond energy correction and linear regression model, respectively (see text for details). Error metrics given are mean unsigned error (MUE) and mean signed error (MSE) with their standard deviations in parentheses, linear correlation coefficient (LCC), and sum of the squares of the deviations (S).

Model		CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
OH								
Raw	MUE	0.9 (*0.7*)	1.1 (*0.8*)	1.2 (*0.8*)	1.0 (*0.9*)	1.0 (*0.9*)	1.7 (*0.9*)	1.7 (*1.0*)
	MSE	-0.1 (*1.2*)	0.5 (*1.3*)	0.7 (*1.2*)	0.1 (*1.4*)	0.4 (*1.3*)	-1.6 (*1.2*)	-1.6 (*1.2*)
	S	24.381	31.889	34.482	33.285	31.089	67.804	68.687
BEC	MUE	0.9 (*0.7*)	1.0 (*0.7*)	1.0 (*0.7*)	1.1 (*0.9*)	1.0 (*0.8*)	0.8 (*0.8*)	0.8 (*0.8*)
	MSE	-0.0 (*1.2*)	-0.0 (*1.3*)	-0.0 (*1.2*)	0.0 (*1.4*)	0.0 (*1.3*)	0.0 (*1.2*)	0.0 (*1.2*)
	S	24.274	27.729	25.808	33.196	27.920	23.501	23.501
LRM	MUE	0.6 (*0.6*)	0.7 (*0.6*)	0.8 (*0.5*)	0.9 (*0.8*)	0.8 (*0.7*)	0.8 (*0.8*)	0.9 (*0.8*)
	MSE	-0.0 (*0.8*)	0.0 (*0.9*)	-0.0 (*1.0*)	-0.0 (*1.3*)	0.0 (*1.1*)	0.0 (*1.1*)	0.0 (*1.2*)
	S	11.656	14.891	15.529	26.839	20.982	21.951	23.206
SH								
RAW	MUE	0.7 (*0.3*)	0.2 (*0.2*)	0.3 (*0.1*)	0.7 (*0.4*)	1.3 (*0.5*)	0.3 (*0.2*)	0.3 (*0.2*)
	MSE	-0.7 (*0.3*)	-0.2 (*0.3*)	0.1 (*0.3*)	0.7 (*0.4*)	1.3 (*0.5*)	0.0 (*0.4*)	-0.0 (*0.4*)
	S	3.074	0.447	0.435	3.357	8.983	0.736	0.746
BEC	MUE	0.3 (*0.1*)	0.2 (*0.1*)	0.2 (*0.1*)	0.3 (*0.2*)	0.3 (*0.3*)	0.3 (*0.2*)	0.3 (*0.2*)
	MSE	-0.0 (*0.3*)	-0.0 (*0.3*)	-0.0 (*0.3*)	0.0 (*0.4*)	0.0 (*0.5*)	-0.0 (*0.4*)	0.0 (*0.4*)
	S	0.388	0.317	0.332	0.626	0.812	0.734	0.745
LRM	MUE	0.3 (*0.1*)	0.2 (*0.1*)	0.2 (*0.1*)	0.3 (*0.3*)	0.3 (*0.3*)	0.3 (*0.2*)	0.3 (*0.2*)
	MSE	0.0 (*0.3*)	0.0 (*0.3*)	-0.0 (*0.3*)	0.0 (*0.4*)	-0.0 (*0.4*)	-0.0 (*0.4*)	-0.0 (*0.4*)
	S	0.386	0.316	0.331	0.608	0.685	0.709	0.722
NH								
Raw	MUE	0.3 (*0.4*)	0.6 (*0.5*)	0.5 (*0.5*)	1.4 (*0.9*)	1.2 (*1.0*)	0.9 (*1.0*)	0.9 (*1.0*)
	MSE	0.1 (*0.5*)	0.6 (*0.5*)	0.4 (*0.5*)	1.4 (*0.9*)	1.2 (*1.0*)	0.5 (*1.2*)	0.5 (*1.2*)
	S	2.280	6.047	4.266	27.750	22.455	16.292	16.149
BEC	MUE	0.3 (*0.4*)	0.3 (*0.4*)	0.3 (*0.4*)	0.7 (*0.6*)	0.7 (*0.6*)	1.0 (*0.7*)	1.0 (*0.7*)
	MSE	0.0 (*0.5*)	0.0 (*0.5*)	0.0 (*0.5*)	-0.0 (*0.9*)	0.0 (*1.0*)	-0.0 (*1.2*)	0.0 (*1.2*)
	S	2.249	2.491	2.439	7.869	8.989	13.840	13.888
LRM	MUE	0.3 (*0.3*)	0.3 (*0.3*)	0.3 (*0.3*)	0.6 (*0.4*)	0.6 (*0.4*)	0.6 (*0.3*)	0.6 (*0.3*)
	MSE	0.0 (*0.5*)	-0.0 (*0.5*)	0.0 (*0.5*)	-0.0 (*0.7*)	-0.0 (*0.7*)	-0.0 (*0.7*)	-0.0 (*0.7*)
	S	2.095	2.000	1.876	4.655	4.687	4.191	4.374

GPB. Δ H_X^C and ΔG_X^C are assigned to enforce the MSE to be zero. This bond energy correction (BEC) model is now compared to a more flexible model based on a linear regression using

$$PA_{pred} = b_X \cdot PA_{model} + a_X \tag{2.9}$$

where PA_{model} is the proton affinity predicted by a computational model, a_X and b_X are parameters optimized to best fit the experimental data, X is either OH, SH, or NH bond, and PA_{pred} is the linear regression model (LRM) prediction. The equation for the GPB is similar to equation 2.9. Parameters for both the BEC and LRM models are provided in supporting information.

Table 2.3 presents PA error metrics for the uncorrected (Raw), BEC, and LRM models. Shown are the MUE and MSE with standard deviation following in parenthesis and the total residual, S (i.e. sum of the deviation squares). Also calculated was the linear correlation coefficients, but for all models this value was near unity. Similar trends are found for the GPB and the table is provided in supporting information.

For the uncorrected (Raw) values, the CBS model outperforms all other models for O-H and N-H compounds as shown by a lower residual. Further, CBS shows a tighter distribution around the experimental values, shown in the standard deviation of the MUE and MSE. The BEC model significantly lowers the residual and MUE for the DFT methods as well as the CBS model for the S-H compounds, as discussed above this is an indication of a systematic error in the methods.

Figure 2.2 displays a histographic comparison of the CBS (the most accurate on average) and QCRNA (the most computationally inexpensive) method for the uncorrected, BEC, and LRM. From this comparison it is apparent how the uncorrected QCRNA is systematically underestimating the PA values. After the BEC or LRM correction, the QCRNA has a more evenly distributed error distribution, but is still is not as tightly distributed as the CBS model.

Summarizing the model comparison, all models are generally reliable within 1-3 kcal/mol. The CBS model generally is the most reliable model except for the S-H compounds, which is easily overcome by the BEC or LRM correction. The following sections provide the uncorrected PA and GPB for biomolecules, but these corrections are easily applied through the coefficients provided in the supporting information and Equations 2.8 and 2.9.

2.3.2 Amino Acids

Amino acids contain two protonation sites along the backbone (an amine and a carboxylic acid) and possibly one or two more on the side chain. In the gas-phase, the amino acid backbone does not exist in isolation as a zwitterion as it does in solution.[141,142] It is known that the backbone titratable sites, isolated in gas-phase, are highly sensitive to the side chain conformation.[95,143] Here we report model compounds for the amino acid side chains to reduce the conformational complexity[144] and focus more directly on the intrinsic affinity of each site for protonation. For example acetic acid is used for aspartic acid as interactions with various backbone conformations would bias the intrinsic affinity of the side chain protonation site. These model compounds are based on the amino acid side chain capped by a methyl group. Table 2.4 reports the PA and GPB for the amino acid model compounds as well as both titratable sites of glycine and proline to provide reference values for the backbone sites.

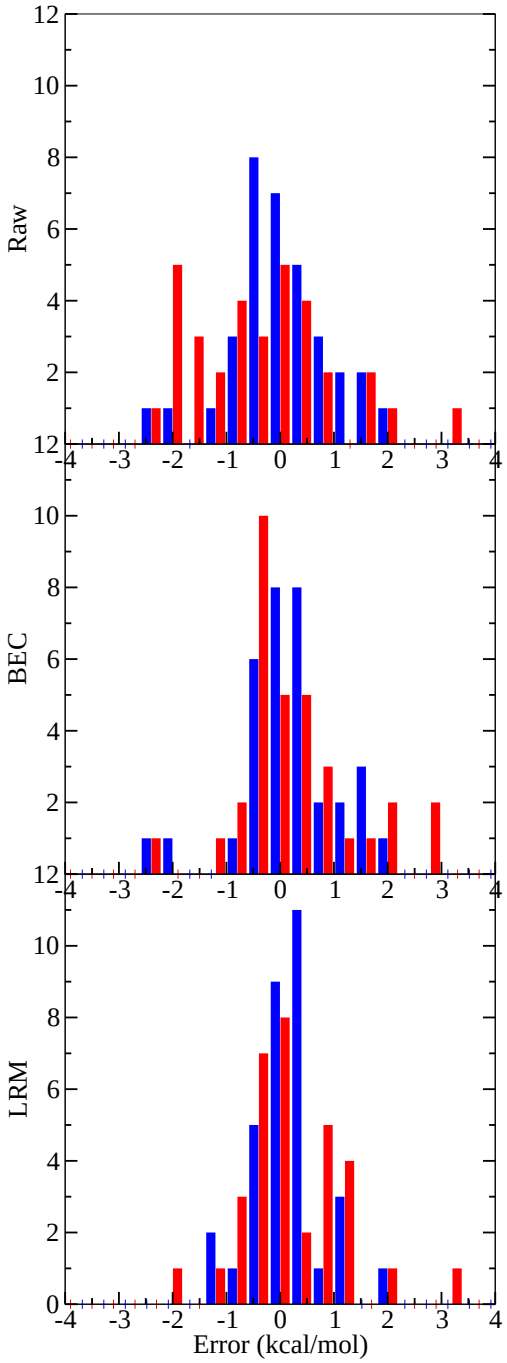

Figure 2.2: Proton affinity error histogram for the CBS (blue,left bar) and QCRNA (red,right bar) against experiment (top), with bond energy correction (middle), and linear regression model (bottom). See text for bond energy correction (BEC) and linear regression model (LRM) information. Histogram bins are width 0.5 kcal/mol.

Table 2.4: Proton affinities and gas-phase basicities for amino acid model compounds with experimental value[a] (experimental error in parentheses) and pK_a values[b]. All quantities are in kcal/mol except pK_a.

Model (Amino Acid)	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA	Experiment	pK_a
Proton Affinity									
prolinium (Pro)	224.3	224.8	225.0	225.4	225.3	225.0	225.0	224.0 (*2.0*)	1.9
glycinium (Gly)	211.9	212.3	212.2	213.6	213.2	212.0	212.0	211.9 (*0.5*)	2.4
acetic acid (Asp)	347.4	348.2	348.6	347.6	347.9	345.8	346.4	347.2 (*1.1*)	3.9
propanic acid (Glu)	346.2	347.0	347.4	347.8	347.9	346.1	345.5	347.4 (*1.8*)	4.1
4-methyl-imidazolium (His π)	229.0	229.6	229.4	231.3	231.2	230.7	230.7	227.7 (*2.0*)	6.0
4-methyl-imidazolium (His τ)	229.6	230.3	230.1	231.8	231.7	231.3	231.3	- (-)	6.0
methanethiol (Cys)	356.8	357.4	357.7	358.2	358.7	357.3	357.3	357.3 (*1.2*)	8.4
glycine (Gly)	341.8	342.6	343.0	343.0	343.2	341.5	341.5	340.3 (*1.1*)	9.8
p-methylphenol (Tyr)	350.4	350.7	350.7	350.0	350.6	348.9	348.9	350.7 (*1.3*)	10.5
propanammonium (Lys)	219.3	219.8	219.6	220.3	220.1	219.4	219.3	219.4 (*0.5*)	10.5
proline (Pro)	340.1	341.0	341.2	342.1	342.1	340.4	340.4	340.3 (*3.1*)	10.6
n-methylguanidine (Arg)	239.5	239.6	239.0	243.1	242.9	241.9	241.8	- (-)	12.5
Gas Phase Basicity									
prolinium (Pro)	216.8	217.3	217.4	217.8	217.6	217.4	217.4	216.0 (*2.0*)	1.9
glycinium (Gly)	204.1	204.5	204.4	205.4	205.2	204.0	204.2	203.7 (*0.5*)	2.4
acetic acid (Asp)	339.4	340.2	340.6	341.3	341.5	339.5	338.3	341.4 (*1.2*)	3.9
propanic acid (Glu)	340.1	341.0	341.4	339.2	339.9	337.6	339.4	340.4 (*1.4*)	4.1
4-methyl-imidazolium (His π)	221.2	221.8	221.6	223.6	223.5	223.0	223.0	220.1 (*2.0*)	6.0
4-methyl-imidazolium (His τ)	221.9	222.6	222.4	224.2	224.0	223.7	223.6	- (-)	6.0
methanethiol (Cys)	350.4	351.0	351.3	351.8	352.3	351.0	350.9	350.6 (*2.0*)	8.4
glycine (Gly)	334.7	335.6	335.9	335.8	336.0	334.2	334.3	335.1 (*1.4*)	9.8
p-methylphenol (Tyr)	343.2	343.4	343.4	342.8	343.2	341.8	341.6	343.8 (*1.2*)	10.5
propanammonium (Lys)	211.8	212.3	212.2	212.8	212.6	211.8	211.9	211.3 (*0.5*)	10.5
proline (Pro)	332.5	333.4	333.6	334.4	334.4	332.6	332.6	333.4 (*3.0*)	10.6
n-methylguanidine (Arg)	232.7	232.9	232.3	236.8	237.3	235.5	235.0	- (-)	12.5

[a]All experimental values are from NIST (Ref. 121), except for prolinium which can be found in reference 145. [b] pK_a values are taken from reference 146.

Overall, the quantum models are self consistent and reliably predict the experimental values for the model compounds. The largest errors are for histidine, though most models are still within the experimental error. For arginine (n-methyl-guanidine) no experimental value was found. Hunter and Lias report a PA of 235.7 kcal/mol and GPB of 226.9 kcal/mol for guanidine,[147] which is close to our prediction. Norberg *et al.* report a computational prediction for n-methyl-guanidine PA as 242 - 248 kcal/mol (MP2/6-31G*), but notes that this structure is basis set dependent and that correlated methods provide lower values. Our quantum models diverge for this compound, with the DFT models predicting a significantly larger (∼3 kcal/mol) PA and GPB compared to the multi-level models, consistent with Norberg *et al.* suggestion.

Histidine values are provided both for the π (near backbone) and τ (far from backbone) nitrogen positions,[148] sometimes referred to as δ and σ, respectively. All the models show a preference for deprotonation at the π position, but only by ∼0.5 kcal/mol, which indicates the system environment and solvation will dominate the site preference.

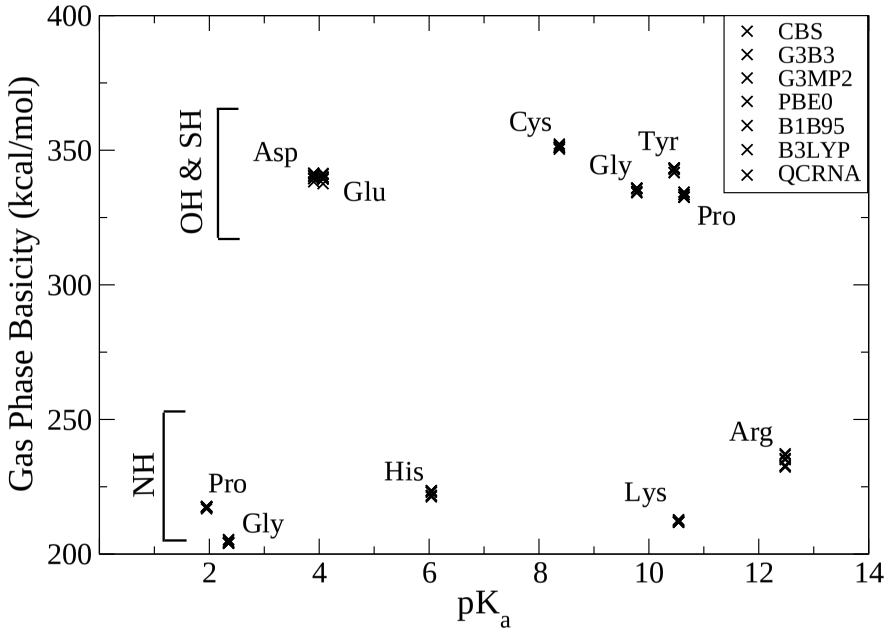

Figure 2.3: Calculated gas phase basicities of amino acid model compounds versus pK_{a} values.

It is noteworthy that while certain residues have very disparate pK_a values they can have similar GPB and alternatively similar pK_a values can have significantly different GPB. For example, tyrosine and glutamic acid only differ 2 kcal/mol (equivalent to 1.4 pK_a units) but are 6.4 pK_a units apart, indicating most of the difference is solvation. Alternatively, glycinium's GPB is 12.3 kcal/mol (9 pK_a units) greater than prolinium, but only 0.5 pK_a units.

Figure 2.3 shows the correlation between the GPB and pK_a for the amino acids for each computational model. It is apparent that the PA and GPB do not have a linear correlation to the pK_a values. Because the pK_a and GPB (as in Figure 2.1) are related by

$$-\mathrm{GPB} = \Delta \mathrm{G_{gas}} = \frac{\mathrm{RT}}{\log(\mathrm{e})} p\mathrm{K_a} + \mathrm{constant} \tag{2.10}$$

where the constant is a combination of solvation energies. Therefore, if the solvation energy differences are close, one would expect to see a slope of -1.364 kcal/mol/pK_a unit at 298.15 K This explains the general grouping shown in the figure as the O-H and S-H compounds have different charges than the N-H compounds(anions vs. cations) leading to significantly different solvation. The O-H amino acids show a slope of -0.448, close to the ideal value, while the N-H compounds have a slope of 1.019, the inverse of what is expected. This difference makes sense in the context of how relatively dissimilar he solvation of the N-H compounds are expected to be (e.g. arginine compared with glycinium).

2.3.3 Nucleic Acid Bases

The DNA and RNA bases each contain multiple protonation sites, but the most biologically relevant are shown in Figure 2.4[149,150] along with their pK_a values. In addition, these bases have tautomeric forms, which are accessible in some biological systems.[151] A complete list of the PA and GPB of all protonation sites for DNA/RNA bases in both their keto and enol form as well as the tautomerization energies are provided in supporting information. Table 2.5 gives the PA and GPB errors compared to experimental values provided by Lias and Hunter[147] and pK_a values[151] for the protonation sites in Figure 2.4.

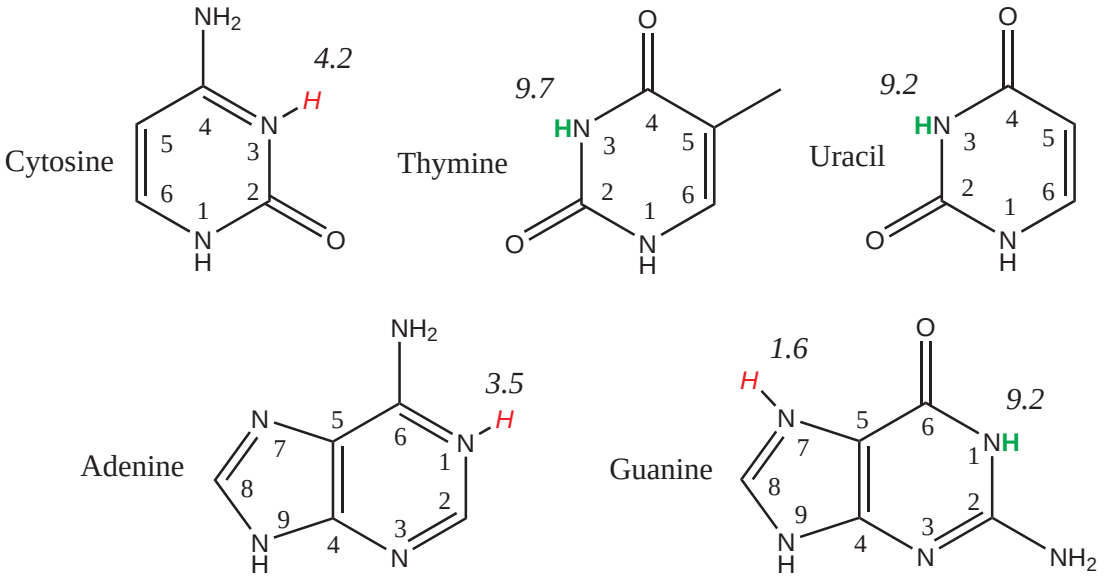

Figure 2.4: DNA/RNA basepair protonation points with pK_a values. Hydrogens in green are points of deprotonation. Hydrogens in red are points of protonation.

Compared to the experimental data and amino acids considered above, these errors for both the PA and GPB are larger. In particular, thymine and uracil have some of the largest errors for the multi-level models, even after proton orientation and conformational averaging were considered. Additionally, CBS seems to dramatically underestimate the values. Alternatively if we compare our computation to experimental values provided by Greco *et al.* for Ade (223.5 kcal/mol), Cyt (223.8 kcal/mol), and Thy (209.8 kcal/mol) we see better correlation.[152] Wolken and Tureček make a similar argument a PA for 205.1 kcal/mol for uracil.[153] This may also explain difficulties using the Lias data when interpreting ammonia proton transfer with nucleic acid bases.[154]

The bases fall well within the expected trend between GPB and pK_a, with a slope of -2.61 and correlation coefficient of -0.9797. This is likely because of the similar solvation each base has upon protonation (all nitrogen positions). We show that the PA and GPB follow the trend of

$$\text{GUA} > \text{CYT} > \text{ADE} > \text{THY} > \text{URA}$$

which is consistent with previous studies.[155,156]

Table 2.5: Proton affinities (top) and gas-phase basicities (bottom) errors (calculated - experimental value) for DNA and RNA bases. The uncertainty of the experimental values is ≈ 2.0 kcal/mol.[147] All quantities are in kcal/mol except pK_a.

Base	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA	Expt.	pK_a
Proton Affinity									
Ade	-1.7	-0.4	0.6	1.8	1.8	1.4	1.5	225.3	3.5
Cyt	-0.1	0.6	0.4	2.5	2.3	1.9	1.8	227.0	4.2
Gua	-1.7	-1.1	-1.4	1.7	1.5	1.2	1.2	229.3	1.6
Thy	-4.4	-3.4	-3.2	-1.3	-1.2	-2.4	-2.3	210.5	9.7
Ura	-4.2	-3.2	-3.1	-1.3	-1.2	-2.2	-2.2	208.6	9.2
Gas Phase Basicity									
Ade	-1.8	-0.9	-1.1	0.2	0.3	0.1	0.2	218.1	3.5
Cyt	-0.1	0.8	0.7	1.9	1.2	1.4	0.9	219.0	4.2
Gua	-1.5	-0.8	-1.1	1.9	1.6	1.3	1.3	221.7	1.6
Thy	-4.8	-3.9	-3.7	-1.8	-1.7	-2.8	-2.7	203.2	9.7
Ura	-4.5	-3.6	-3.4	-1.6	-1.6	-2.6	-2.5	201.2	9.2

2.3.4 Ribose

One of the defining differences between DNA and RNA is the RNA 2' hydroxyl group. This group is a very weak acid, but plays an important role in DNA and RNA cleavage and ligation[47] Based on experimental and computation estimates, the O2' pK_a is between 12.5 and 14.9.[108,157,158] This range is likely due to the local chemical environment, sugar pucker, and solution ionic strength.

Full quantum structural optimization leads the sugar puckers outside what is considered biologically relevant. To provide the most useful values comparable to the other biological residues, we have calculated the PA and GPB using C2'-endo (P=163.797, τ_m=34.495) and C3'-endo (P=13.506, τ_m=37.747) sugar puckers.[159] These conformations are based on x-ray diffraction[160,161] and are consistent with B and A form DNA and RNA.[151,162] Additionally, dihedral synonymous to the the ϵ dihedral (C4'-C3'-O2'-P) for phosphate backbone was in the biological trans conformation.

Similar to the amino acids, use of the full ribose structure created conformational obstacles that obfuscate the intrinsic acidity of the 2' oxygen. In particular a hydroxy group at the 3' position created a significant intramolecular hydrogen bond that biased the O2' values. To get the most representative values, we calculated three model compounds

Table 2.6: Proton Affinities (top) and gas-phase basicities (bottom) errors (calculated - experimental value) for RNA like sugar molecules: 3-hydroxy tetrahydrofuran (THF), 3-hydroxy-2,4,5-methyl-tetrahydrofuran (Methyl), and 3-hydroxy-2,5-methyl-4-methoxy-tetrahydrofuran (Ribose). All molecules are given in conformations consistent with B-form and A-form DNA/RNA. All quantities are in kcal/mol.

Pucker	Model	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
				Proton Affinity				
C2'[a]	THF	369.2	369.1	370.1	370.2	370.6	368.2	368.3
	Methyl	366.4	367.0	367.4	367.3	367.4	365.6	365.7
	Ribose	360.4	361.2	361.6	362.9	363.2	360.8	361.5
C3'[b]	THF	370.2	370.9	371.2	370.8	371.3	368.7	368.8
	Methyl	367.7	368.3	368.7	368.2	368.6	366.4	366.5
	Ribose	364.9	365.7	366.1	365.6	365.8	363.9	363.9
				Gas Phase Basicity				
C2'[a]	THF	361.2	360.9	362.0	362.1	362.4	360.1	360.1
	Methyl	358.0	358.6	359.0	358.7	358.8	357.1	357.2
	Ribose	354.5	355.4	355.8	354.9	355.1	355.1	353.5
C3'[b]	THF	362.9	363.5	363.9	363.6	364.1	361.4	361.5
	Methyl	360.3	360.9	361.3	360.8	361.2	359.0	359.2
	Ribose	357.8	358.6	359.0	358.5	358.7	356.8	356.8

[a]P=163.797, τ_m4=34.495. [b]P=13.506,τ_m=37.747.

for ribose: 3-hydroxy-tetrahydrofuran (THF), 2-hydroxy-2,4,5-trimethyl-tetrahydrofurn (Methyl), and 3-hydroxy-2,5-dimethyl-4-methoxy-tetrahydrofuran (Ribose). These successively more complicated models give some insight into how the ribose structure tunes the intrinsic PA and GPB of the 2'-hydroxyl group, with Ribose being the most similar to the sugar ring of RNA.

Table 2.6 provides the PA and GPB for THF, Methyl, and Ribose. Unlike the nucleic acid bases, the multi-level and DFT models behavior similar to the experimental comparison, indicating CBS will be the most reliable indicator of PA and GPB. C3'-endo has a greater PA and GPB for all models compared to the C2'-endo sugar pucker conformation, and this gap increases with model complexity. Taking CBS as the most reliable model for an alcohol, the GPB is 3.3 kcal/mol larger and more acidic for C3'-endo. The addition of the 3' methyl then the 3' methoxy group successively raises the PA and GPB of the 2' hydroxyl group making it less likely to release its proton. This is likely due, in particular for ribose, due to the stabilization of the protonated structure via intramolecular hydrogen bonding, as shown by the difference with the THF model and the larger effect for C2'-endo conformation.

2.3.5 Phosphates

The protonation state of phosphoranes is important and has implications for phosphate transesterification/hydrolysis mechanism.[163] Of particular interest, is the affect on the transition states encountered in ribozyme biochemistry.[103,164,165] Knowing the PA and GPB can also be used as an indication of hydrogen bonding potential, which has implication for nucleic acid structure.[166]

To fully investigate the protonation states of the phosphate moiety, we provide values for metaphosphates, phosphates, cyclic phosphates, phosphoranes, and cyclic phosphoranes, each with thio-substitutions. Figure 2.5 provides the nomenclature for phosphate structures. Table 2.7 reports the PA for the metaphosphates, phosphates, and phosphoranes using the CBS and QCRNA methods. PA and GPB shows similar trends and values are reported using all the quantum models and can be found in the supporting information.

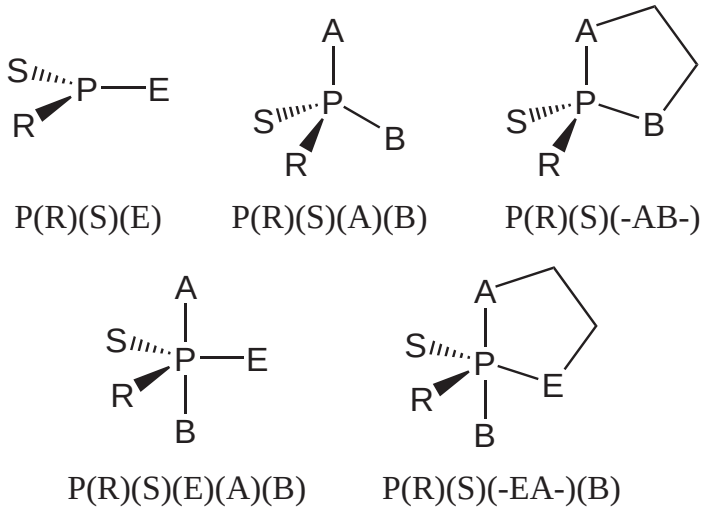

Figure 2.5: Nomenclature convention for ligand designations in metaphosphate, acyclic and cyclic phosphate and phosphorane compounds of biological interest. This nomenclature is consistent with the naming convention used for similar compounds in previous work.[97, 130, 139]

Table 2.7: Predicted proton affinities for metaphosphate, phosphate, and cyclic phosphate (left columns) and acyclic and cyclic phosphorane compounds (right column) of biological interest. All values are in kcal/mol. See Figure 2.5 for naming scheme.

Molecule[a]	CBS	QCRNA	Molecule[a]	CBS	QCRNA
P(O)(O)(OH)	310.5	311.1	P(OH*)(OH)(OH)(OH)(OH)	340.8	339.3
P(O)(O)(SH)	304.5	306.2	P(OH)(OH)(OH)(OH)(OH*)	350.8	350.5
P(S)(O)(OH)	307.5	308.5	P(OH*)(OH)(-O-CH_2CH_2-O-)(OH)	343.1	342.6
P(S)(S)(OH)	307.2	308.6	P(OH)(OH*)(-O-CH_2CH_2-O-)(OH)	343.4	342.9
P(S)(O)(SH)	303.3	305.2	P(OH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	343.8	343.0
P(O)(OH)(OH)(OH)	327.9	328.2	P(OH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	343.4	342.9
P(O)(O)(OH)(OH) $^-$	458.7	457.8	P(OH*)(OH)(-O-CH_2CH_2-S-)(OH)	335.2	335.0
P(O)(O)(O)(OH) $^{2-}$	580.9	579.5	P(OH)(OH*)(-O-CH_2CH_2-S-)(OH)	335.4	335.4
P(S)(OH)(OH)(OH)	322.5	322.9	P(OH*)(OH)(-O-CH_2CH_2-S-)(OCH_3)	331.8	334.5
P(O)(OH)(OH)(SH*)	318.2	319.9	P(OH)(OH*)(-O-CH_2CH_2-S-)(OCH_3)	332.1	334.4
P(O)(OH)(OH*)(SH)	320.9	320.9	P(OH*)(OH)(-S-CH_2CH_2-O-)(OH)	336.1	336.2
P(O)(OCH_3)(OH)(OH)	330.0	330.3	P(OH)(OH*)(-S-CH_2CH_2-O-)(OH)	336.0	336.2
P(O)(OCH_3)(O)(OH) $^-$	453.9	452.9	P(OH*)(OH)(-S-CH_2CH_2-O-)(OCH_3)	330.0	338.6
P(S)(OCH_3)(OH)(OH)	323.5	324.1	P(OH)(OH*)(-S-CH_2CH_2-O-)(OCH_3)	339.6	338.8
P(S)(OCH_3)(O)(OH) $^-$	437.9	437.5	P(OH*)(OH)(-O-CH_2CH_2-O-)(SCH_3)	333.1	333.0
P(O)(OCH_3)(OH)(SH*)	319.1	321.4	P(OH)(OH*)(-O-CH_2CH_2-O-)(SCH_3)	333.1	333.0
P(O)(OCH_3)(OH*)(SH)	321.6	322.1	P(SH*)(OH)(-O-CH_2CH_2-O-)(OH)	333.6	335.0
P(O)(SCH_3)(OH)(OH)	322.2	322.8	P(SH)(OH*)(-O-CH_2CH_2-O-)(OH)	338.4	337.7
P(O)(SCH_3)(O)(OH) $^-$	443.7	441.0	P(SH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	335.3	336.8
P(O)(OCH_3)(OCH_3)(OH)	329.2	330.1	P(SH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	339.1	338.2
P(S)(OCH_3)(OCH_3)(OH)	325.1	325.9	P(OH*)(SH)(-O-CH_2CH_2-O-)(OH)	337.3	336.7
P(O)(SCH_3)(OCH_3)(OH)	324.2	326.2	P(OH)(SH*)(-O-CH_2CH_2-O-)(OH)	332.6	334.1
P(O)(OCH_3)(OCH_3)(SH)	320.8	323.2	P(OH*)(SH)(-O-CH_2CH_2-O-)(OCH_3)	341.1	340.1
P(O)(OH)(-O-CH_2CH_2-O-)	329.4	329.5	P(OH)(SH*)(-O-CH_2CH_2-O-)(OCH_3)	335.9	336.8
P(O)(SH)(-O-CH_2CH_2-O-)	320.1	321.9			
P(S)(OH)(-O-CH_2CH_2-O-)	324.0	324.5			
P(O)(OH)(-S-CH_2CH_2-O-)	324.1	324.7			

[a] "Molecule" refers to AH in Eq. 2.1

Metaphosphate has long been a subject of investigation because of its roll in phosphoryl transfer reaction[167–169] and in particular for ribozyme catalysis[170,171] where the metaphosphate is formed during the dissociative mechanism.[172] We report a PA and GPB for metaphosphate of 310.5 kcal/mol 304.2 kcal/mol, respectively, using the CBS method. These values 20-80 kcal/mol lower than that of most other experimentally listed O-H compounds above.

Phosphates, both acyclic and cyclic and methylated, have PA between 327.9 - 329.4 kcal/mol. Given the differences in structure, it is interesting how small a range is predicted. Also reported are the three deprotonations PA of phosphoric acid. Given the challenges in calculating pK_{a} values due to the difficulty in describing the solvation of charged species, predicting multiple absolute pK_{a} values within 1 pK_{a} unit is generally out of reach of computational methods. This benchmark data can be used in conjunction with the experimentally known pK_{a} values to help guide future solvation work.

The cyclic phosphoranes are an analogue to the transition state or intermediate during phosphate transesterification, hydrolysis, or ligation.[173] For example, during RNA backbone cleavage the deprotonated O2' attacks the phosphate forming a penta-coordinated phosphate followed by the bond breakage of the O5' oxygen.[47] This attacking group is synonymous to atom A, the leaving group with atom B, and the non-bridging oxygens with S and R, and the O3' oxygen to atom E in Figure 2.5. During the concerted mechanism in the catalyzed reaction, the negative charge that builds up along the non-bridging oxygens must be neutralized. The PA and GPB at these positions may guide mechanistic interpretations. In particular, is that the pro-R and pro-S non-bridging oxygens are not equivalent in many biological systems. We predict the pro-R and pro-S non-bridging oxygen PAs and GPBs differ by ≈0.5 kcal/mol. This is primarily due to the 5-membered ring structures pucker (P-A-C-C-E) and the leaving group orientation (B). Further the anionic structures were only stable in the gas-phase when the remaining non-bridging proton was closest to the leaving group (B). If the proton points along the P-A bond, the structure cleaves along this bond leaving what would be the starting product of ribozyme backbone transesterification.

Thio-substitution is useful biochemical technique when investigating reaction mechanism about phosphates.[174–177] For all phosphates and phosphoranes thio-substition at the deprotonation point lowers the PA by ≈8-10 kcal/mol making them less likely to

less likely to donate the proton. For the non-bridging positions of phosphates the affect is ≈4-8 kcal/mol, while phosphoranes range from ≈4-7 kcal/mol non-bridging oxygen. Because the dianionic penta-coordinate phosphate is unstable in the gas-phase, sulfurs in the non-bridging non-ring position are protonated, likely shifting the values up. In the case of deprotonated sulfur at the non-bridging positions (both pro-S and pro-R), the actual affect on the PA of the other oxygen will likely be toward the lower end of that range if the trend of the phosphates holds.

2.4 Conclusion

Benchmark calculations of the proton affinity and gas-phase basicity for oxygen, sulfur, and nitrogen containing are compared for a variety of multi-level and density functional quantum methods against experimental values. These methods are corrected using a bond energy correction and linear regression model to evaluate the quantum methods' performance. Here we report that while all the tested quantum chemistries provide reliable predictions, the CBS-QB3 method in general is the most reliable.

These models are then used to predict the proton affinity and gas-phase basicity for approximately 150 different molecules of biological interest comprising more than 2000 quantum calculations. These systems include amino acid backbone and side chains, nucleic acids in two tautomeric forms, ribose like structures for RNA sugar's O2', and various metaphosphates, cyclic and acyclic phosphates, and phosphoranes all with and without thio-substitution.

This work provides a consistent database of PA and GPB to be used in pK_{a} prediction, biochemical evaluation of protonation/deprotonation events, parameterization of future semiempirical quantum models used in QM/MM biocatalysis simulation, and comparison to experimental PA and GPB determination.

2.5 Supporting Information

Table 2.8: Coefficients for proton affinity (top) and gas phase basicity (bottom) bond energy correction (BED) and linear regression (LRM) fit models for experimental data. See text for parameter details. ΔH^C and ΔG^C for the BEC model and b for the LRM model are in kcal/mol. The a parameter for the LRM model is unitless.

		Multi-level methods			DFT methods			
Model		CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Proton Affinity								
BEC	ΔH^C_{O}	0.1	-0.5	-0.7	-0.1	-0.4	1.6	1.6
LRM	a_{OH}	15.2	14.8	13.1	10.8	11.0	7.0	7.2
LRM	b_{OH}	0.957	0.957	0.961	0.969	0.968	0.984	0.984
BEC	ΔH^C_{S}	0.7	0.2	-0.1	-0.7	-1.3	-0.0	0.0
LRM	a_{SH}	-3.3	3.0	1.5	10.0	25.4	12.3	11.9
LRM	b_{SH}	1.011	0.992	0.995	0.970	0.925	0.965	0.967
BEC	ΔH^C_{N}	-0.1	-0.6	-0.4	-1.4	-1.2	-0.5	-0.5
LRM	a_{NH}	3.8	6.2	6.2	15.4	18.0	26.9	26.8
LRM	b_{NH}	0.982	0.969	0.969	0.924	0.913	0.875	0.876
Gas Phase Basicity								
BEC	ΔG^C_{O}	0.5	-0.1	-0.3	0.4	0.0	2.0	2.0
LRM	a_{OH}	16.4	16.1	14.4	12.3	12.2	8.8	9.0
LRM	b_{OH}	0.954	0.953	0.958	0.966	0.965	0.980	0.980
BEC	ΔG^C_{S}	0.9	0.3	0.0	-0.5	-1.1	0.2	0.2
LRM	a_{SH}	27.7	32.2	29.4	30.5	40.1	31.7	28.1
LRM	b_{SH}	0.923	0.908	0.916	0.911	0.882	0.909	0.920
BEC	ΔG^C_{N}	-0.2	-0.8	-0.6	-1.5	-1.3	-0.6	-0.7
LRM	a_{NH}	0.5	3.0	3.0	12.4	14.6	23.0	22.6
LRM	b_{NH}	0.996	0.982	0.982	0.934	0.925	0.889	0.890

Table 2.9: Gas-phase basicity error analysis for each model broken down by bond type and fitting method. Raw indicates no correction has been made. BEC and LRM are bond energy correction and linear regression model, respectively (see text for details). Error metrics given are mean unsigned error (MUE) and mean signed error (MSE) with their standard deviations in parentheses, linear correlation coefficient (LCC), and sum of the squares of the deviations (S).

Model		CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
OH								
Raw	MUE	1.0 (*0.7*)	0.9 (*0.7*)	0.9 (*0.6*)	1.0 (*0.8*)	0.8 (*0.8*)	2.0 (*1.0*)	2.0 (*1.0*)
	MSE	-0.5 (*1.1*)	0.1 (*1.1*)	0.3 (*1.1*)	-0.4 (*1.3*)	-0.0 (*1.1*)	-2.0 (*1.0*)	-2.0 (*1.0*)
	S	24.898	22.527	21.875	29.516	20.541	90.719	91.721
BEC	MUE	0.8 (*0.7*)	0.9 (*0.6*)	0.9 (*0.5*)	0.9 (*0.8*)	0.8 (*0.8*)	0.7 (*0.7*)	0.7 (*0.7*)
	MSE	-0.0 (*1.1*)	-0.0 (*1.1*)	-0.0 (*1.1*)	-0.0 (*1.3*)	-0.0 (*1.1*)	-0.0 (*1.0*)	-0.0 (*1.0*)
	S	20.466	22.458	20.517	26.696	20.537	16.693	16.693
LRM	MUE	0.4 (*0.4*)	0.5 (*0.4*)	0.6 (*0.4*)	0.9 (*0.5*)	0.7 (*0.5*)	0.7 (*0.5*)	0.8 (*0.5*)
	MSE	0.0 (*0.6*)	0.0 (*0.7*)	-0.0 (*0.7*)	-0.0 (*1.0*)	-0.0 (*0.8*)	0.0 (*0.9*)	0.0 (*1.0*)
	S	5.823	7.446	8.253	18.731	12.196	14.159	15.812
SH								
RAW	MUE	0.9 (*0.5*)	0.5 (*0.3*)	0.4 (*0.2*)	0.6 (*0.4*)	1.1 (*0.4*)	0.4 (*0.3*)	0.3 (*0.3*)
	MSE	-0.9 (*0.5*)	-0.3 (*0.5*)	-0.0 (*0.5*)	0.5 (*0.5*)	1.1 (*0.4*)	-0.2 (*0.4*)	-0.2 (*0.4*)
	S	5.189	1.613	0.971	2.319	6.743	0.998	0.867
BEC	MUE	0.4 (*0.2*)	0.4 (*0.2*)	0.4 (*0.2*)	0.3 (*0.3*)	0.3 (*0.2*)	0.3 (*0.3*)	0.3 (*0.2*)
	MSE	-0.0 (*0.5*)	0.0 (*0.5*)	-0.0 (*0.5*)	0.0 (*0.5*)	-0.0 (*0.4*)	0.0 (*0.4*)	0.0 (*0.4*)
	S	1.161	1.072	0.970	0.819	0.689	0.800	0.586
LRM	MUE	0.4 (*0.2*)	0.4 (*0.2*)	0.4 (*0.2*)	0.3 (*0.2*)	0.2 (*0.2*)	0.3 (*0.2*)	0.2 (*0.2*)
	MSE	-0.0 (*0.5*)	0.0 (*0.5*)	0.0 (*0.5*)	0.0 (*0.4*)	0.0 (*0.3*)	-0.0 (*0.4*)	-0.0 (*0.3*)
	S	1.043	0.898	0.825	0.654	0.375	0.628	0.453
NH								
Raw	MUE	0.5 (*0.3*)	0.8 (*0.5*)	0.7 (*0.4*)	1.5 (*0.8*)	1.3 (*0.9*)	1.0 (*0.8*)	1.0 (*0.8*)
	MSE	0.2 (*0.5*)	0.8 (*0.5*)	0.6 (*0.6*)	1.5 (*0.8*)	1.3 (*0.9*)	0.6 (*1.1*)	0.7 (*1.1*)
	S	3.152	8.852	6.917	29.538	23.652	14.708	14.808
BEC	MUE	0.4 (*0.4*)	0.4 (*0.3*)	0.4 (*0.4*)	0.5 (*0.6*)	0.5 (*0.7*)	0.7 (*0.8*)	0.7 (*0.8*)
	MSE	-0.0 (*0.5*)	0.0 (*0.5*)	0.0 (*0.6*)	-0.0 (*0.8*)	-0.0 (*0.9*)	0.0 (*1.1*)	-0.0 (*1.1*)
	S	2.592	2.375	2.878	5.940	6.797	10.580	10.418
LRM	MUE	0.4 (*0.4*)	0.4 (*0.3*)	0.4 (*0.4*)	0.4 (*0.4*)	0.5 (*0.4*)	0.4 (*0.3*)	0.4 (*0.3*)
	MSE	0.0 (*0.5*)	0.0 (*0.5*)	0.0 (*0.5*)	-0.0 (*0.6*)	0.0 (*0.6*)	-0.0 (*0.5*)	-0.0 (*0.5*)
	S	2.586	2.207	2.649	3.437	3.462	2.561	2.645

Table 2.10: Predicted proton affinities for DNA and RNA bases in keto/amino tautomeric form.

Base	Protonation	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Adenine	N1 (+)	223.6	224.9	224.7	227.1	227.1	226.7	226.8
	N3 (+)	222.0	223.3	223.1	225.8	225.7	225.5	225.5
	N6 (+)	202.5	203.5	203.7	202.9	203.0	202.6	202.6
	N6 (−)	355.3	355.0	354.4	357.3	357.5	355.9	355.9
	N7 (+)	215.3	216.5	216.4	219.1	219.0	218.8	218.9
Cytosine	O2 (+)	228.4	229.1	229.1	230.0	230.2	228.9	229.0
	N3 (+)	226.9	227.6	227.4	229.5	229.3	228.9	228.8
	N4 (+)	196.0	196.6	196.9	195.8	195.9	195.4	195.3
	N4 (−)	347.7	348.1	347.6	350.8	350.9	349.4	349.3
Guanine	N1 (−)	337.7	338.2	338.0	340.9	341.0	339.9	339.9
	N2 (+)	190.1	190.4	190.7	190.1	190.2	190.0	190.0
	N2 (−)	338.7	339.1	338.8	340.3	340.5	339.1	339.0
	N3 (+)	211.4	212.1	211.9	214.0	214.0	213.8	213.9
	O6 (+)	222.4	223.3	223.2	225.5	225.5	224.6	224.7
	N7 (+)	227.6	228.2	227.9	231.0	230.8	230.5	230.5
Thymine	O2 (+)	200.2	201.1	201.4	201.6	201.8	201.0	201.1
	N3 (+)	177.3	178.1	178.6	177.6	177.6	178.6	178.5
	N3 (−)	346.3	347.0	346.9	348.2	348.1	347.1	346.9
	O4 (+)	206.1	207.1	207.3	209.2	209.3	208.1	208.2
Uracil	O2 (+)	197.0	197.7	198.0	197.7	198.0	197.2	197.2
	N3 (+)	174.4	175.9	176.3	175.3	175.4	176.4	176.3
	N3 (−)	345.9	346.6	346.5	347.7	347.8	346.5	346.4
	O4 (+)	204.4	205.4	205.5	207.3	207.4	206.4	206.4

Table 2.11: Predicted gas-phase basicity for DNA and RNA bases in keto/amino tautomeric form.

Base	Protonation	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Adenine	N1 (+)	216.3	217.1	217.0	218.3	218.4	218.2	218.3
	N3 (+)	214.7	215.5	215.4	217.0	217.0	216.9	217.1
	N6 (+)	196.0	196.5	196.7	194.9	195.1	194.8	195.0
	N6 (−)	347.5	347.7	347.0	351.1	351.1	349.6	349.3
	N7 (+)	208.5	209.4	209.2	210.7	210.7	210.6	210.8
Cytosine	O2 (+)	220.2	221.1	221.1	221.2	220.9	220.3	219.8
	N3 (+)	218.9	219.8	219.7	220.9	220.2	220.4	219.9
	N4 (+)	187.8	188.5	188.8	186.8	186.5	186.7	186.1
	N4 (−)	340.1	339.8	339.3	344.0	344.6	342.3	342.4
Guanine	N1 (−)	330.2	330.6	330.4	333.3	333.5	332.4	332.3
	N2 (+)	182.8	183.2	183.5	182.8	182.9	182.7	182.8
	N2 (−)	331.2	331.5	331.2	332.8	333.1	331.5	331.4
	N3 (+)	204.6	205.4	205.3	207.1	207.1	206.9	207.1
	O6 (+)	215.0	216.1	216.0	218.1	218.1	217.2	217.3
	N7 (+)	220.2	220.9	220.6	223.6	223.3	223.0	223.0
Thymine	O2 (+)	192.7	193.7	194.0	194.1	194.4	193.6	193.7
	N3 (+)	169.4	171.8	172.2	171.1	171.0	172.4	172.2
	N3 (−)	338.5	339.1	339.1	340.4	340.4	339.3	339.0
	O4 (+)	198.4	199.3	199.5	201.4	201.5	200.4	200.5
Uracil	O2 (+)	189.1	190.2	190.5	190.3	190.5	189.7	189.8
	N3 (+)	168.1	169.5	169.9	168.7	168.8	170.1	169.9
	N3 (−)	338.1	338.6	338.5	339.9	340.0	338.7	338.4
	O4 (+)	196.7	197.6	197.8	199.6	199.6	198.6	198.7

Table 2.12: Enthalpy (top) and free energy (bottom) of tautomerization for DNA and RNA basepairs. All quantities are in kcal/mol.

	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Proton Affinity							
Adenine	12.5	11.7	11.4	12.3	12.2	11.9	11.9
Cytosine	1.2	1.0	0.8	2.4	2.4	2.1	2.1
Guanine	-0.4	-0.2	-0.3	0.7	0.5	1.1	1.1
Thymine	12.0	11.8	11.6	12.3	11.9	12.5	12.5
Uracil	11.1	10.9	10.7	11.3	10.9	11.6	11.5
Gas Phase Basicity							
Adenine	12.3	12.0	11.7	13.6	13.4	13.0	12.8
Cytosine	1.5	1.2	1.0	3.4	3.8	3.0	3.4
Guanine	-0.3	0.0	-0.1	0.8	0.6	1.2	1.2
Thymine	12.2	12.0	11.8	12.5	12.0	12.7	12.6
Uracil	11.2	11.1	10.8	11.4	11.1	11.7	11.6

Table 2.13: Predicted proton affinities for DNA and RNA bases in enol/imino tautomeric form.

Base	Protonation	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Adenine	N1 (−)	342.8	343.3	343.0	345.0	345.2	344.0	344.0
	N3 (+)	212.1	212.7	212.6	214.1	214.1	214.0	214.0
	N6 (+)	236.1	236.5	236.1	239.4	239.3	238.7	238.6
	N6 (−)[a]	353.3	358.1	357.7	365.0	362.7	357.0	357.2
	N7 (+)	218.6	224.4	219.0	221.4	221.3	221.1	221.1
Cytosine	O2 (+)	203.0	203.8	204.0	203.7	204.1	203.2	203.2
	N3 (−)	346.5	347.0	346.8	348.4	348.5	347.2	347.2
	N4 (+)	228.1	228.6	228.2	231.9	231.7	231.0	231.0
	N4 (−)[a]	364.4	364.1	363.7	371.4	368.6	363.6	363.7
Guanine	N1 (+)	222.0	223.2	223.0	226.2	226.0	225.7	225.8
	N2 (+)	206.5	207.1	207.3	207.4	207.4	206.7	206.7
	N2 (−)	358.1	357.9	357.4	360.1	360.3	358.4	358.3
	N3 (+)	217.6	218.6	218.4	222.0	221.9	221.5	221.6
	O6 (−)	338.1	338.4	338.3	340.2	340.6	338.8	338.8
	N7 (+)	224.7	225.7	225.5	228.6	228.5	228.3	228.3
Thymine	O2 (+)	221.6	222.4	222.5	224.0	224.0	222.9	222.9
	N3 (+)	218.0	218.9	218.9	221.5	221.2	220.6	220.6
	O4 (−)	334.4	335.2	335.3	335.9	336.2	334.5	334.5
Uracil	O2 (+)	217.9	218.6	218.8	220.1	220.1	218.9	219.0
	N3 (+)	215.5	216.3	216.2	218.6	218.3	217.9	217.9
	O4 (−)	334.9	335.6	335.8	336.4	336.8	335.0	334.9

[a] Pyrimidine or pyridine ring breaks.

Table 2.14: Predicted gas-phase basicity for DNA and RNA bases in enol/imino tautomeric form.

Base	Protonation	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
Adenine	N1 (−)	335.0	335.5	335.2	337.3	337.5	336.3	336.2
	N3 (+)	204.7	205.3	205.2	206.7	206.7	206.6	206.7
	N6 (+)	228.6	229.1	228.7	231.9	231.8	231.2	231.2
	N6 (−)[a]	375.6	349.2	348.9	356.2	353.8	348.2	348.3
	N7 (+)	211.0	211.6	211.4	213.8	213.7	213.5	213.5
Cytosine	O2 (+)	195.6	196.3	196.5	196.3	196.6	195.7	195.8
	N3 (−)	343.7	343.8	343.4	345.4	345.5	344.2	343.6
	N4 (+)	220.5	221.0	220.6	224.3	224.1	223.4	223.3
	N4 (−)[a]	354.0	353.7	353.2	361.0	358.1	353.1	353.2
Guanine	N1 (+)	214.7	216.2	216.0	218.9	218.7	218.4	218.5
	N2 (+)	199.4	200.0	200.3	200.3	200.3	199.6	199.5
	N2 (−)	350.4	350.1	349.6	352.4	352.6	350.6	350.5
	N3 (+)	210.3	211.5	211.4	214.7	214.5	214.2	214.3
	O6 (−)	330.4	330.6	330.5	332.5	332.9	331.1	331.1
	N7 (+)	217.1	218.2	218.0	221.0	220.8	220.6	220.6
Thymine	O2 (+)	213.8	214.6	214.8	216.3	216.3	215.1	215.1
	N3 (+)	210.5	211.3	211.3	213.9	213.5	213.1	213.1
	O4 (−)	326.4	327.1	327.3	328.0	328.4	326.6	326.4
Uracil	O2 (+)	210.1	210.8	210.9	212.3	212.3	211.1	211.2
	N3 (+)	207.9	208.7	208.6	211.0	210.7	210.3	210.3
	O4 (−)	326.9	327.5	327.7	328.5	328.9	327.0	326.8

[a] Pyrimidine or pyridine ring breaks.

Table 2.15: Predicted proton affinities for metaphosphate, phosphate, and cyclic phosphate compounds of biological interest.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
P(O)(O)(OH)	310.5	311.0	311.8	312.4	312.5	311.0	311.1
P(O)(O)(SH)	304.5	305.2	305.7	306.2	306.5	306.2	306.2
P(S)(O)(OH)	307.5	308.0	308.7	309.4	310.0	308.4	308.5
P(S)(S)(OH)	307.2	307.9	308.5	309.3	310.2	308.5	308.6
P(S)(O)(SH)	303.3	303.8	304.4	304.8	305.5	305.2	305.2
P(O)(OH)(OH)(OH)	327.9	328.3	328.7	330.0	329.9	328.1	328.2
$P(O)(O)(OH)(OH)^-$	458.7	458.6	459.4	460.5	460.8	457.8	457.8
$P(O)(O)(O)(OH)^{2-}$	580.9	578.3	580.9	583.1	583.8	579.4	579.5
P(S)(OH)(OH)(OH)	322.5	322.9	323.4	324.5	324.8	322.9	322.9
$P(O)(OH)(OH)(SH^*)$	318.2	318.8	319.0	320.2	320.4	320.1	319.9
$P(O)(OH)(OH^*)(SH)$	320.9	321.2	321.9	322.7	322.7	320.9	320.9
$P(O)(OCH_3)(OH)(OH)$	330.0	330.6	331.0	331.8	331.8	330.3	330.3
$P(O)(OCH_3)(O)(OH^*)$	453.9	453.8	454.6	455.2	455.5	452.8	452.9
$P(S)(OCH_3)(OH)(OH)$	323.5	324.0	324.6	325.3	325.6	324.0	324.1
$P(S)(OCH_3)(O)(OH)^-$	437.9	438.2	438.9	439.7	440.2	437.4	437.5
$P(O)(OCH_3)(OH)(SH^*)$	319.1	319.9	320.1	321.1	321.5	321.4	321.4
$P(O)(OCH_3)(OH^*)(SH)$	321.6	322.0	322.6	323.5	323.4	322.1	322.1
$P(O)(SCH_3)(OH)(OH)$	322.2	322.4	323.0	324.3	324.4	322.8	322.8
$P(O)(SCH_3)(O)(OH)^-$	443.7	444.2	444.4	444.8	444.9	441.6	441.0
$P(O)(OCH_3)(OCH_3)(OH)$	329.2	329.9	330.3	331.0	330.9	330.0	330.1
$P(S)(OCH_3)(OCH_3)(OH)$	325.1	325.8	326.3	326.7	327.0	325.8	325.9
$P(O)(SCH_3)(OCH_3)(OH)$	324.2	326.1	326.7	327.4	327.4	326.1	326.2
$P(O)(OCH_3)(OCH_3)(SH)$	320.8	321.7	321.9	322.4	322.7	323.2	323.2
$P(O)(OH)(\text{-O-}CH_2CH_2\text{-O-})$	329.4	329.9	330.4	330.3	330.5	329.4	329.5
$P(O)(SH)(\text{-O-}CH_2CH_2\text{-O-})$	320.1	321.0	321.2	321.7	323.9	321.9	321.9
$P(S)(OH)(\text{-O-}CH_2CH_2\text{-O-})$	324.0	324.6	325.2	325.6	330.9	324.4	324.5
$P(O)(OH)(\text{-S-}CH_2CH_2\text{-O-})$	324.1	324.5	325.1	326.1	326.2	324.7	324.7

[a] "Molecule" refers to the neutral molecule AH in Eq. 2.1

Table 2.16: Predicted proton affinities for phosphorane compounds of biological interest.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
P(OH*)(OH)(OH)(OH)(OH)	340.8	340.6	341.3	341.8	341.8	339.2	339.3
P(OH)(OH)(OH)(OH)(OH*)	350.8	351.2	351.5	352.4	352.5	350.6	350.5
P(OH*)(OH)(-O-CH_2CH_2-O-)(OH)	343.1	343.6	344.2	344.4	344.6	342.4	342.6
P(OH)(OH*)(-O-CH_2CH_2-O-)(OH)	343.4	344.0	344.5	344.8	344.9	342.8	342.9
P(OH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	343.8	344.6	345.1	344.8	345.0	342.9	343.0
P(OH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	343.4	344.3	344.7	344.5	344.6	342.7	342.9
P(OH*)(OH)(-O-CH_2CH_2-S-)(OH)	335.2	335.9	336.4	337.4	337.5	334.9	335.0
P(OH)(OH*)(-O-CH_2CH_2-S-)(OH)	335.4	336.1	336.7	337.8	337.8	335.3	335.4
P(OH*)(OH)(-O-CH_2CH_2-S-)(OCH_3)	331.8	335.1	335.6	336.5	336.3	334.3	334.5
P(OH)(OH*)(-O-CH_2CH_2-S-)(OCH_3)	332.1	335.5	335.9	336.7	336.8	334.1	334.4
P(OH*)(OH)(-S-CH_2CH_2-O-)(OH)	336.1	336.3	336.9	338.3	338.3	336.2	336.2
P(OH)(OH*)(-S-CH_2CH_2-O-)(OH)	336.0	336.2	336.8	338.2	338.2	336.1	336.2
P(OH*)(OH)(-S-CH_2CH_2-O-)(OCH_3)	330.0	339.5	340.1	340.4	340.5	338.5	338.6
P(OH)(OH*)(-S-CH_2CH_2-O-)(OCH_3)	339.6	340.0	340.7	340.8	341.0	338.8	338.8
P(OH*)(OH)(-O-CH_2CH_2-O-)(SCH_3)	333.1	333.8	334.2	335.0	335.1	332.8	333.0
P(OH)(OH*)(-O-CH_2CH_2-O-)(SCH_3)	333.1	333.8	334.2	335.0	335.1	332.8	333.0
P(SH*)(OH)(-O-CH_2CH_2-O-)(OH)	333.6	334.4	334.6	335.0	335.4	335.0	335.0
P(SH)(OH*)(-O-CH_2CH_2-O-)(OH)	338.4	338.9	339.5	339.5	339.7	337.5	337.7
P(SH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	335.3	336.4	336.3	336.6	337.0	336.8	336.8
P(SH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	339.1	339.7	340.3	339.9	340.2	338.1	338.2
P(OH*)(SH)(-O-CH_2CH_2-O-)(OH)	337.3	337.7	338.4	338.6	338.8	336.6	336.7
P(OH)(SH*)(-O-CH_2CH_2-O-)(OH)	332.6	333.3	333.5	334.1	334.5	334.1	334.1
P(OH*)(SH)(-O-CH_2CH_2-O-)(OCH_3)	341.1	341.7	342.1	341.9	342.1	339.9	340.1
P(OH)(SH*)(-O-CH_2CH_2-O-)(OCH_3)	335.9	336.9	337.1	337.0	337.7	336.8	336.8

[a] "Molecule" refers to the neutral molecule AH in Eq. 2.1

Table 2.17: Predicted gas-phase basicities for metaphosphate, phosphate, and cyclic phosphate compounds of biological interest.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
P(O)(O)(OH)	304.2	304.8	305.6	306.1	306.2	304.7	304.8
P(O)(O)(SH)	297.8	298.5	299.1	299.5	299.7	299.5	299.5
P(S)(O)(OH)	300.5	301.1	301.9	302.4	303.0	301.4	301.6
P(S)(S)(OH)	300.2	301.1	301.7	302.3	303.2	301.5	301.7
P(S)(O)(SH)	296.6	297.2	297.8	298.1	298.8	298.5	298.5
P(O)(OH)(OH)(OH)	320.7	321.0	321.5	322.9	322.8	321.0	321.0
$P(O)(O)(OH)(OH)^-$	451.1	451.3	452.1	452.9	453.1	450.1	450.0
$P(O)(O)(O)(OH)^{2-}$	575.4	572.7	575.3	577.5	578.2	573.8	574.1
P(S)(OH)(OH)(OH)	314.4	315.0	315.5	316.9	317.2	315.2	314.9
$P(O)(OH)(OH)(SH^*)$	311.3	311.9	312.1	314.2	313.9	313.5	313.0
$P(O)(OH)(OH^*)(SH)$	313.7	314.3	314.9	316.4	315.9	314.0	313.9
$P(O)(OCH_3)(OH)(OH)$	322.8	323.3	323.7	324.7	324.6	323.2	323.1
$P(O)(OCH_3)(O)(OH^*)$	447.0	447.1	447.8	448.3	448.8	445.9	446.1
$P(S)(OCH_3)(OH)(OH)$	316.6	317.3	317.8	318.4	318.8	317.1	317.2
$P(S)(OCH_3)(O)(OH)^-$	430.8	431.1	431.9	432.6	433.2	430.3	430.5
$P(O)(OCH_3)(OH)(SH^*)$	313.2	313.6	313.8	314.6	314.8	314.9	314.5
$P(O)(OCH_3)(OH^*)(SH)$	315.3	315.5	316.1	316.7	316.6	315.3	315.2
$P(O)(SCH_3)(OH)(OH)$	315.2	315.5	316.1	317.5	317.4	315.9	315.9
$P(O)(SCH_3)(O)(OH)^-$	434.5	435.1	435.3	436.7	436.8	433.1	434.2
$P(O)(OCH_3)(OCH_3)(OH)$	322.9	323.4	323.9	324.8	324.9	323.8	323.7
$P(S)(OCH_3)(OCH_3)(OH)$	318.0	318.7	319.2	319.9	320.2	318.9	318.7
$P(O)(SCH_3)(OCH_3)(OH)$	317.7	318.8	319.4	319.7	319.6	318.5	318.9
$P(O)(OCH_3)(OCH_3)(SH)$	314.6	315.2	315.4	316.0	316.3	316.9	316.9
$P(O)(OH)(\text{-}O\text{-}CH_2CH_2\text{-}O\text{-})$	321.2	322.3	322.9	324.3	324.6	321.4	321.8
$P(O)(SH)(\text{-}O\text{-}CH_2CH_2\text{-}O\text{-})$	313.0	313.9	314.1	314.4	316.5	314.7	314.7
$P(S)(OH)(\text{-}O\text{-}CH_2CH_2\text{-}O\text{-})$	316.7	317.4	317.9	318.1	323.3	316.9	317.1
$P(O)(OH)(\text{-}S\text{-}CH_2CH_2\text{-}O\text{-})$	317.0	317.3	317.9	319.1	319.3	317.5	317.5

[a] "Molecule" refers to the neutral molecule AH in Eq. 2.1

Table 2.18: Predicted gas-phase basicities for phosphorane compounds of biological interest.

Molecule[a]	CBS	G3B3	G3MP	PBE0	B1B95	B3LYP	QCRNA
P(OH*)(OH)(OH)(OH)(OH)	333.7	333.2	333.9	334.4	334.5	331.6	331.7
P(OH)(OH)(OH)(OH)(OH*)	343.4	343.8	344.0	345.0	345.0	343.0	342.5
P(OH*)(OH)(-O-CH_2CH_2-O-)(OH)	335.8	336.5	337.1	337.2	337.4	335.1	335.4
P(OH)(OH*)(-O-CH_2CH_2-O-)(OH)	336.0	336.8	337.3	337.4	337.6	335.4	335.6
P(OH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	336.0	336.8	337.3	337.0	337.2	335.0	335.1
P(OH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	335.6	336.5	337.0	336.7	336.8	334.9	335.0
P(OH*)(OH)(-O-CH_2CH_2-S-)(OH)	327.2	328.0	328.6	329.8	329.9	326.9	327.2
P(OH)(OH*)(-O-CH_2CH_2-S-)(OH)	327.5	328.4	329.0	330.2	330.1	327.3	327.7
P(OH*)(OH)(-O-CH_2CH_2-S-)(OCH_3)	323.8	327.2	327.7	328.7	328.5	326.4	326.7
P(OH)(OH*)(-O-CH_2CH_2-S-)(OCH_3)	323.7	327.3	327.7	328.5	328.5	325.7	326.1
P(OH*)(OH)(-S-CH_2CH_2-O-)(OH)	328.8	328.9	329.6	331.1	331.2	329.0	328.8
P(OH)(OH*)(-S-CH_2CH_2-O-)(OH)	328.6	328.8	329.5	330.9	331.0	328.8	328.7
P(OH*)(OH)(-S-CH_2CH_2-O-)(OCH_3)	331.0	331.5	332.1	332.5	333.0	330.5	330.5
P(OH)(OH*)(-S-CH_2CH_2-O-)(OCH_3)	331.5	331.9	332.5	332.8	333.2	330.7	330.7
P(OH*)(OH)(-O-CH_2CH_2-O-)(SCH_3)	325.3	326.0	326.4	327.4	327.5	325.1	325.1
P(OH)(OH*)(-O-CH_2CH_2-O-)(SCH_3)	325.3	326.0	326.4	327.4	327.5	325.1	325.1
P(SH*)(OH)(-O-CH_2CH_2-O-)(OH)	326.5	327.3	327.5	327.9	328.4	327.9	327.8
P(SH)(OH*)(-O-CH_2CH_2-O-)(OH)	330.5	331.0	331.6	331.9	332.2	329.8	329.9
P(SH*)(OH)(-O-CH_2CH_2-O-)(OCH_3)	327.9	329.1	329.0	329.3	329.7	329.5	329.3
P(SH)(OH*)(-O-CH_2CH_2-O-)(OCH_3)	331.3	332.0	332.6	332.1	332.4	330.2	330.4
P(OH*)(SH)(-O-CH_2CH_2-O-)(OH)	329.5	329.9	330.6	331.0	331.3	329.0	329.0
P(OH)(SH*)(-O-CH_2CH_2-O-)(OH)	325.4	326.2	326.4	327.1	327.5	327.1	326.8
P(OH*)(SH)(-O-CH_2CH_2-O-)(OCH_3)	332.9	333.5	334.0	334.0	334.2	331.9	332.1
P(OH)(SH*)(-O-CH_2CH_2-O-)(OCH_3)	328.2	329.3	329.4	329.4	330.0	329.1	329.2

[a] "Molecule" refers to the neutral molecule AH in Eq. 2.1

Chapter 3

Influence of C5 cytosine substitution in base pairs with guanine

The work in this chapter is the product of a collaboration between members of the York Group (Adam Moser and Professor Darrin M. York) and the Tretyakova Group (Rebecca Guza and Professor Natalia Tretyakova) and is published under the title "Density Functional Study of the Influence of C5 Cytosine Substitution in Base Pairs with Guanine".[120]

3.1 Introduction

Methylation at the C5 position of cytosine (^{Me}C) is an important endogenous nucleobase modification in mammalian genomic DNA[178–180] found in approximately 1% of total bases in mammalian genome.[181] This epigenetic modification, which is a conserved change to DNA,[182] formed by specific methyltransferase enzymes,[180] influences chromatin structure[183–185] and mediates gene expression[186] and occurs in 60% to 90% of vertebrate CG dinucleotides.[186] Many tumors exhibit altered methylation patterns, leading to activation of proto-oncogenes and silencing of tumor suppressor genes.[187] Endogenously methylated CG within the coding region of the *p53* tumor suppressor

gene[188] , particularly codons 157, 158, 245, 248, and 273,[188–191] are known lung cancer mutational hotspots.[192,193] The presence of mutational hotspots at methylated CG dinucleotides can be caused by an increased chemical reactivity toward carcinogens. If not repaired prior to DNA replication, lesions at the coding region can cause polymerase errors and induce heritable mutations.[194]

It has been shown that ^{Me}C can modulate the reactivity of neighboring guanine towards carcinogens and DNA alkylating drugs[195–204] and its influence depends on the attacking species. Guanine reactivity towards the tobacco carcinogen benzo[a]pyrene,[205] which is metabolized into the reactive polycyclic aromatic hydrocarbon (+)-*anti*-7*r*,8*t*-dihyrdoxy-*c*9,10-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene (BPDE), is doubled upon cytosine methylation.[206] Another prominent tobacco carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), exhibits reduced reactivity at methylated CG dinucleotides.[204,207] This implies that C5 cytosine methylation modulates reactivity differently depending on the reactive species.

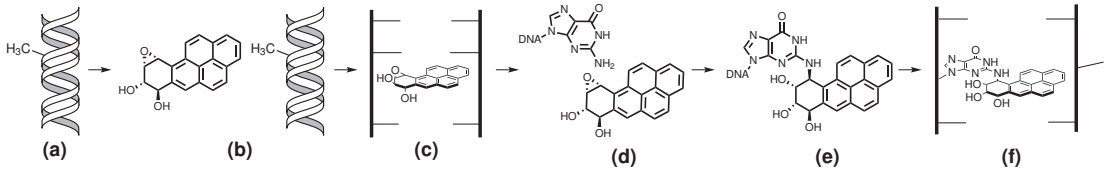

Figure 3.1: General steps involved in BPDE reaction with DNA containing 5-methyl cytosine. (a) DNA duplex containing ^{Me}C in the major groove. (b) BPDE pre-covalently binds at the DNA grooves. (c) BPDE intercalates into the DNA duplex. (d) BPDE reaches a reactive conformation with the exocyclic amine of guanine. (e) BPDE reacts at N^2 of guanine forming a covalent adduct. (f) The adduct reaches a stable final conformation.

While the ability of ^{Me}C to influence the reactivity of neighboring guanine is well established, the question of how such a relatively small chemical modification on cytosine can influence specific adduct yields at the neighboring guanine bases remains unanswered. As illustrated in Fig. 3.1, C5 methylation can influence the susceptibility of neighboring guanine bases toward carcinogen attack by modifying DNA structure, pre-covalent binding in the major groove, intercalation, and local structure and electronics at the ^{Me}CG step. Further, methylation is known to affect the conformations of

the resulting DNA adducts, which may play a role in repair.[208,209] To paint a complete mechanistic picture that will provide insight into carcinogenesis, all these steps must be considered.

One strategy uniquely suited to elucidate this multi-step process is to compare C5 cytosine analogs to ^{Me}C . In a study using mitomycin C, fluorine was used in place of the methyl substituent.[199] This electron withdrawing analog gave insight into the importance of electronic effects on reactivity. Another study showed how C5 halogen substituted cytosine and uracil can mimic methylation and affect ^{Me}C signals observed in tumors.[210] C5 cytosine analogs have also been studied computationally.[211,212] However, one cannot unambiguously interpret the effects of a single, or even a few, chemical modifications. In fact, the complexity of the general mechanism of DNA adduct formation, as depicted in Fig. 3.1 with BPDE, requires a very broad range of analogs that exhibit distinct, systematic chemical trends is required to make any definitive statements about ^{Me}C role in individual reaction steps.

Theoretical methods provide a powerful tool to aid in the interpretation of experimental data and to describe the effects of chemical modifications, including their effect on local electronic structure properties of individual C5 substituted bases,[211] base pairs,[56,119,212–214] and their more global effect on the helical base stack.[215–218] These factors will influence the biologically relevant chemical reactivity. In the present work, we undertake the study of a systematic series of cytosine analogs shown in Fig. 3.2, with the goal of providing insight into the electronic structure properties of the C5 modified cytosine bases and their base pairs with protonated and unprotonated guanine. Modifications include alkyl, alkenyl and alkynyl chains, halogens, aromatic, fused ring, and strong σ and π withdrawing and electron donating functional groups. This information will be useful, ultimately, in arriving at a consensus view of methylated cytosine's role in mediating DNA reactivity with carcinogens and drugs.

3.2 Methods

All calculations were performed in accord with the standardized protocol, previously detailed,[88] used to construct the *QCRNA* database, a recently developed on-line database of quantum calculations for RNA catalysis.[219] A summary of the *QCRNA* protocol is

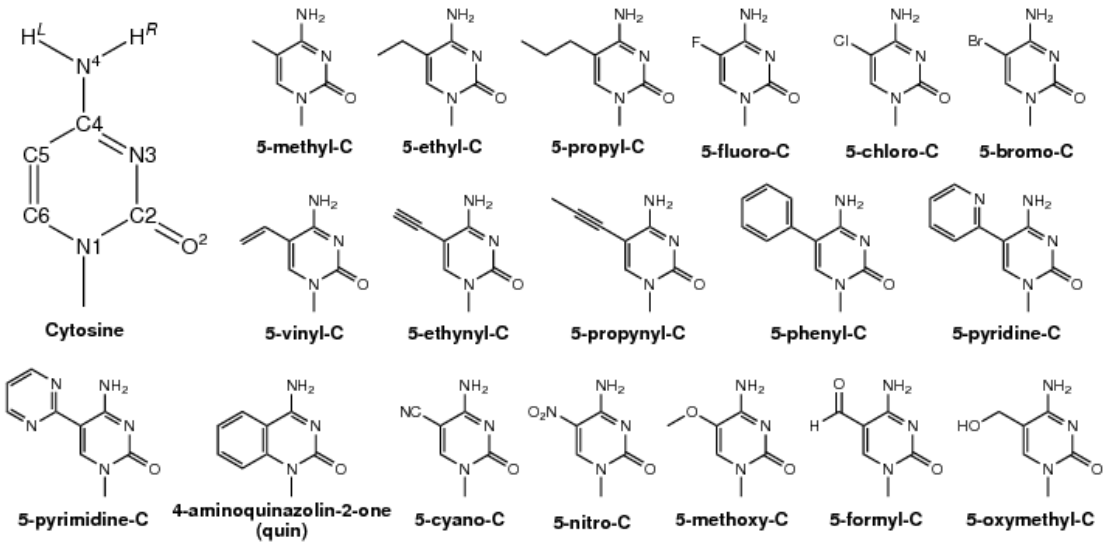

Figure 3.2: Numbered cytosine (top left) along with cytosine analogs with substitution at the 5 position.

as follows:

- B3LYP/6-31++G(d,p) geometry optimization to a stationary point on the adiabatic potential energy surface
- B3LYP/6-31++G(d,p) wave function stability check
- B3LYP/6-31++G(d,p) vibrational frequency analysis (including enthalpic and free energy corrections to the energy at 298.15 K) and polarizability calculation
- B3LYP/6-311++G(3df,2p) gas phase single point energy refinement, evaluation of multipole moments, natural bond orbital (NBO) analysis, and ChelpG atomic charges
- B3LYP/6-311++G(3df,2p) solution-phase single point energy refinement with COSMO solvation, evaluation of multipole moments, natural bond orbital (NBO) analysis, and ChelpG atomic charges
- B3LYP/6-311++G(3df,2p) solution-phase single point energy refinement with PCM solvation, evaluation of multipole moments, natural bond orbital (NBO) analysis, and ChelpG atomic charges

In brief, all structures were optimized in the gas phase with Kohn-Sham density functional theory (DFT) methods using the hybrid exchange functional of Becke[62,63] and the Lee, Yang, and Parr correlation functional[64] (B3LYP) with B3LYP/6-31++G** as implemented in the Gaussian03 suite of programs.[122] Geometry optimizations were performed without constraints in redundant internal coordinates using the 6–31++G(d,p) basis set and the stability of the restricted closed shell Kohn-Sham determinant for each final structure was verified.[220,221] Frequency calculations at the optimized geometries were performed to establish the nature of all stationary points and used to calculate thermodynamic quantities. Electronic energies and other properties of the density, such as moments of the density and natural bond order (NBO) analysis,[222] were further refined via single point calculations at the optimized geometries using the 6–311++G(3df,2p) basis set. All single point calculations were run with convergence criteria on the SCF wave function tightened to 10^{-8} *au* to ensure high precision for properties sensitive to the use of diffuse basis functions.[223] Solvation effects were treated by single-point calculations based on the gas phase optimized structures using the polarizable continuum model (PCM)[77,81,224] and a variation of the conductor-like screening model (COSMO)[225] with the 6-311++G(3df,2p) basis set.

The molecular enthalpies and free energies are based on the refined gas phase single point energy calculations and along with the frequency analysis are used to obtain the proton affinity (PA) and gas phase basicity (GPB). As discussed in previous work,[139,140] $\mathrm{p}K_\mathrm{a}$ values are calculated using the following relationship

$$\Delta G^\circ_{aq} = \frac{RT}{\log(e)} pK_a \tag{3.1}$$

and a standard thermodynamic cycle, shown in Figure 3.3, (e.g. Scheme 1 (a) in reference 226) to give an equation for the $\mathrm{p}K_\mathrm{a}$.

$$pK_a = \frac{\log(e)}{RT}\left[\Delta G^\circ_{\mathrm{gas}} - \Delta G^\circ_{\mathrm{solv}}\left(HA\right) + \Delta G^\circ_{\mathrm{solv}}\left(H^+\right) + \Delta G^\circ_{\mathrm{solv}}\left(A^-\right)\right] \tag{3.2}$$

The solvation free energy of the proton, $\Delta G^\circ_{\mathrm{solv}}\left(H^+\right)$, is -265.87 kcal/mol as determined by Tissandier *et al.*[227,228]

Relative $\mathrm{p}K_\mathrm{a}$ values ($\Delta\mathrm{p}K_\mathrm{a}$) are always given in the following form

$$\Delta\mathrm{p}K_a = \mathrm{p}K_a(\mathrm{analogue}) - \mathrm{p}K_a(\mathrm{native}) \tag{3.3}$$

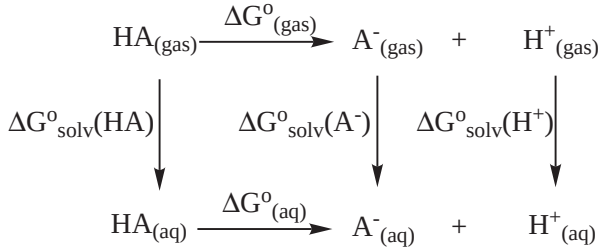

Figure 3.3: Thermodynamic cycle relating the gas-phase basicity, $\triangle G^{\circ}_{gas}$, aqueous reaction free energy, $\triangle G^{\circ}_{aq}$, and the solvation free energies, $\triangle G^{\circ}_{solv}$.

No counterpoise corrections were calculated to correct for basis set superposition errors in the case of the hydrogen bonded base pairs. At the basis set levels used in the present work, these corrections have been shown to be fairly small,[212,229] with differences in relative values typically on the order of 0.1 kcal/mol.

For single base calculations, the N1 proton was substituted with a methyl group to provide a more realistic model of the base connected to a deoxyribose for the single base cytosine calculations.[212] Substituted Cytosine:Guanine base pairs were calculated in their Watson-Crick hydrogen bonding scheme (Fig. 3.4). The same base pairs were calculated with the N^2 of guanine protonated, resulting in a quaternary amino group. All base pair structures retain a coplanar purine and pyridine without constraint.

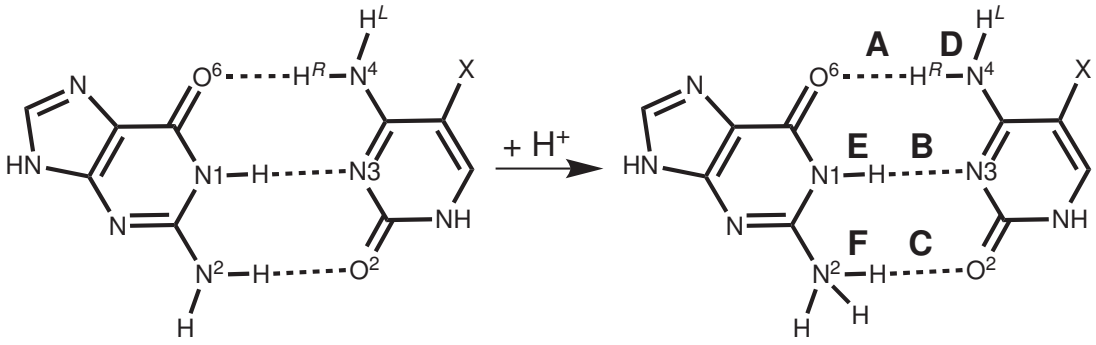

Figure 3.4: Watson-Crick GC base pair and protonated N^2 CG base pair with hydrogen bond lengths labeled.

3.3 Results

The series of cytosine substitutions were chosen as follows. First, consecutively longer alkyl chains were chosen (methyl, ethyl, and propyl) to investigate the possibility that hydrophobic groups may increase carcinogen binding at analog sites, thereby increasing the possibility of reaction at that position. Conversely, there is a possibility larger substituents may begin to impede carcinogens from reaching the reaction site. Secondly, we consider a series of halogens (fluoro, chloro, and bromo). These σ-electron withdrawing substituents can help probe the electronic effects with fairly minimal, systematic steric effects that are complementary to the σ-electron donating alkyls. Thirdly, vinyl, ethynyl, propynyl were added to the list as possible intercalation enhancers. Originally, phenyl was chosen as an aromatic intercalation enhancer, but preliminary calculations indicated the minimum energy conformation does not lie in the plane of the base pair due to steric clash between the H_L amino hydrogen and a nearby hydrogen on the phenyl ring (the in-plane conformer is $\sim$9 kcal/mol higher in free energy). Therefore, fourthly, in addition to phenyl, we considered the set of 2-pyrimidine and 2-pyridine, and 4-aminoquinazolin-2-one (quin) to determine how the relative torsion of these larger aromatic substituents may effect pre-covalent binding and intercalation. Fifthly, we consider strong electron withdrawing and donating groups, including a strong σ-withdrawing cyano, strong π-withdrawing nitro and formyl, and electron donating methoxy and oxymethyl group to further probe electronic effects. All analogs are shown in Fig. 3.2.

The first step in unraveling the effects of these analogs is to consider in detail their effects on individual bases and base pairs. Since cytosine is not covalently bonded to guanine, any electronic or structural effect that influences carcinogen reactivity is likely to occur through the hydrogen bond network, which is known to be cooperative.[230] The results that follow focus on the atoms involved in GC base pair hydrogen bonding. Results are separated into three subsection: cytosine base, cytosine:guanine base pair, and cytosine:guanine base pair with protonated N^2 guanine. All calculations have been deposited in the *QCRNA* database[219] and can be downloaded, viewed, and manipulated on-line.

3.3.1 Cytosine Base

Initially, the effect on cytosine base geometry was considered to determine if substitution makes any significant changes to base structure. Table 3.1 lists the relevant bond lengths, angles, and torsions for atoms potentially involved in hydrogen bonding. Also shown is the substituent length, defined as the distance from C-5 to the farthest substituent atom, which may be an indicator of how well an analog can act as a pre-covalent guide, see Figure 3.1 (b). Bond length and angle changes upon substitution were fairly negligible, less that 0.01 Å and 3 degrees, respectively. The results agree with previously reported structural changes by Dannenberg *et al.*[119] The N^4-C4-H_R-H_L dihedral is an improper dihedral that indicates the degree of pyramidalization at the N^4 amino group, while the H_R-N^4-C4-N3 dihedral indicates the degree to which the H_R proton resides in the plane of the cytosine. These dihedrals may play a role in the strength of base pair hydrogen bonding by helping to align the H_R proton of cytosine to the O^6 of guanine.

Table 3.1: Selected geometric data for cytosine analogues. Values for cytosine bond lengths, angles, and dihedrals are given in Å, degrees, and degrees, respectively.

	Bond Lengths			Angles		Dihedrals		Substituent
Substituent	C2-O^2	N^4-H_L	N^4-H_R	C4-N^4-H_L	C4-N^4-H_R	N^4-C4-H_R-H_L	H_R-N^4-C4-N3	Length[a]
Cytosine	1.226	1.007	1.010	120.9	117.2	11.5	8.7	1.083
Methyl	1.227	1.007	1.010	121.1	116.6	12.6	9.1	2.173
Ethyl	1.228	1.007	1.010	121.1	116.2	13.5	9.5	3.507
Propyl	1.228	1.007	1.010	121.1	116.2	13.2	9.4	4.758
Fluoro	1.226	1.007	1.008	121.4	118.4	0.1	0.1	1.365
Chloro	1.225	1.007	1.009	121.8	118.0	0.1	0.0	1.755
Bromo	1.225	1.007	1.009	121.4	118.1	0.0	0.0	1.899
Vinyl	1.226	1.008	1.011	119.8	115.6	16.6	9.3	3.483
Ethynyl	1.224	1.008	1.008	121.2	118.4	0.1	0.0	3.700
Propynyl	1.226	1.007	1.008	121.0	118.4	0.0	0.0	4.616
Phenyl	1.227	1.008	1.010	120.4	116.2	13.5	9.9	5.395
Pyridine	1.226	1.015	1.009	118.3	116.9	0.6	5.2	5.335
Pyrimidine	1.226	1.014	1.009	119.3	117.4	0.0	0.0	5.264
Quin	1.225	1.006	1.011	121.0	115.4	15.3	9.2	3.910
Cyano	1.221	1.008	1.009	122.0	118.2	0.0	0.0	2.585
Nitro	1.220	1.010	1.009	120.1	117.4	0.0	0.0	2.304
Methoxy	1.229	1.007	1.008	120.7	118.4	0.1	0.1	3.270

In unsubstituted cytosine, the molecular dipole points along the C2-O^2 bond, which is generally true for the analogs though it can be slightly modified. In the CG base pair this dipole would be directed at the N^2 of guanine, so changes in dipole may play a key role in understanding reactivity changes at this position. The magnitude of the dipole moment is significantly reduced for nitro (72%) and cyano (67%) substituents and amplified considerably for methoxy (18%), pyridine (22%), and pyrimidine (22%) substituents. The total isotropic polarizability is increased for all analogs compared to native cytosine, except for fluorine.

Table 3.2: Proton affinity, gas phase basicity, and relative pK_a at various positions of cytosine. Proton affinities and gas phase basicities are reported in kcal/mol. Relative p$K_a$values (ΔpK_a) are relative to the native cytosine, see Eq. 3.3.

	Proton Affinity				Gas Phase Basicity				ΔpK_a^{PCM}		ΔpK_a^{COSMO}	
Substituent	$O^2(+)$	N3(+)	$N_L^4(-)$	$N_R^4(-)$	$O^2(+)$	N3(+)	$N_L^4(-)$	$N_R^4(-)$	$O^2(+)$	N3(+)	$O^2(+)$	N3(+)
Cytosine	231.8	232.4	350.8	355.6	223.6	224.5	342.9	347.5	0.0	0.0	0.0	0.0
Methyl	235.1	234.7	349.7	355.4	227.1	227.0	341.7	347.2	0.9	0.7	1.0	0.7
Ethyl	236.3	235.9	349.5	355.7	228.2	228.1	341.3	346.9	1.1	0.8	1.2	0.9
Propyl	236.8	236.4	349.5	355.7	228.7	228.7	341.3	347.1	1.1	0.8	1.2	0.9
Fluoro	228.6	226.5	347.0	349.5	219.7	217.8	340.1	342.4	-1.5	-2.9	-1.3	-2.9
Chloro	228.5	227.4	346.4	349.4	220.0	219.2	338.9	341.6	-1.7	-2.7	-1.5	-2.5
Bromo	228.8	227.8	345.9	349.0	220.6	219.9	338.0	340.9	-1.4	-2.6	-1.1	-2.5
Vinyl	233.2	233.2	346.8	352.8	225.4	225.8	338.6	345.0	0.0	-0.1	0.2	0.0
Ethynyl	229.4	229.8	348.6	351.1	221.2	221.7	339.8	342.8	-2.4	-2.3	-2.0	0.2
Propynyl	233.7	233.9	351.0	353.6	225.9	226.4	342.9	345.3	-1.0	-1.1	-0.8	-1.0
Phenyl	235.7	236.2	347.6	353.2	227.9	228.8	339.5	345.5	-0.1	-0.1	0.1	0.0
Pyridine	236.6	239.1	346.7	358.2	228.8	231.6	338.2	349.4	-0.5	0.3	-0.4	0.2
Pyrimidine	235.6	238.1	357.4	357.6	227.5	230.2	349.3	348.5	-1.2	-0.4	-1.2	-0.6
Quin	234.9	234.2	342.9	349.5	227.0	226.5	335.6	342.1	0.2	0.2	0.4	0.3
Cyano	219.4	219.8	340.1	343.3	211.3	211.9	332.2	335.2	-4.8	-4.8	-4.6	-4.6
Nitro	217.6	218.9	345.1	343.9	209.7	211.2	336.6	335.5	-5.9	-5.6	-5.5	-5.3
Formyl	223.1	225.9	351.8	350.5	215.2	218.1	342.8	342.1	-4.1	-3.0	-4.1	-3.0
Methoxy	237.0	234.5	352.9	355.0	228.4	226.2	345.9	347.7	0.8	-0.9	0.9	-0.9
Oxymethyl	234.5	235.5	341.0	356.8	226.7	227.9	333.2	347.7	0.4	0.5	0.0	0.2

Table 3.2 provides the proton affinity (PA) and gas phase basicity (GPB) for the four protonation/deprotonation sites: protonation at the O^2 and the N3 and deprotonation of the left (H_L) and right (H_R) N^4 protons (see Fig. 3.2). Analogs show a broad range of effect on the PA and GPB values relative to the native cytosine (∼20 kcal/mol). Trends between analogs are generally conserved between the PA and GPB.

Of particular note are the values for O^2 protonation. According to Dannenberg *et al.*[119] these values can serve as an indicator of how well this oxygen can donate nucleophilicity to the N^2 of base paired guanine, thereby impacting nucleophilic attack. The O^2 PA and GPB increase for alkanes, vinyl, propynyl, aromatic compounds, and methoxy. Of specific interest is the large decreases for both cyano (a σ withdrawer) and nitro (a π withdrawer). The same trend is seen for N3 protonation and given the cooperativity of the hydrogen bonds might also be important to carcinogen reactivity. Similarly, changes to the N^4_R PA and GPB may affect hydrogen bonding to guanine due to its proximity to the substituent. Deprotonation of N^4_L is also shown.

Relative pK_a values are summarized in Table 3.2 for the protonation O^2 and N3, calculated using Eq. 3.2. Neither the PCM nor COSMO model could give reasonable solvation energies for the anionic nucleobases, so pK_a values for the deprotonation at N^4 are not shown. The PCM and COSMO models give very similar results and only slightly different trends for the protonation of O^2 and N3. As a guide to the absolute accuracy of these numbers, Zhang *et al.* cites the cytosine pK_a for the N3 native and methylated analog protonation as 4.45 and 4.6, respectively.[231] Our calculations for the same analogs give 6.29 and 6.98, respectively. Sowers references a 1.8 pK_a drop between 5-fluoro-C and cytosine,[232] while we show a a 2.9 pK_a decrease. While the absolute pK_a difference is ∼2 pK_a units, the relative values that we are relying on are within ∼1.0 pK_a units.

The alkyl, vinyl, propynyl, aromatic, and methoxy analogs all produce minimal pK_a changes for both N3 and O^2 ($\leq$ 1.2 pK_a units). Of the remaining substituents, the pK_a values decrease for both N3 and O^2 by as much as 5.9 pK_a units. These pK_a decreases indicate that the halogen, ethynyl, cyano, and nitro analogs are decreased hydrogen bond acceptors at N3 and O^2 compared to unmodified cytosine.

3.3.2 CG Base Pair

Table 3.3 presents the hydrogen bond lengths and angles for the modified CG base pairs. Similar to the single base calculations, no significant structural deviations are found between the analogs and cytosine. Cyano and nitro analogs show a slight lengthening in the cytosine O^2 - guanine N^2 hydrogen bond along with a one to two degree change in the angle. Overall, the geometrical parameters reported here are close to those reported in other work.[212]

Table 3.3: GC base pair hydrogen bonding geometry. Bond lengths are reported in Å. Angles are reported in degrees.

	Bond Length						Angle		
Substituent	A	D	B	E	C	F	A/D	B/E	C/F
Cytosine	1.759	1.037	1.914	1.034	1.914	1.022	179.2	177.1	178.3
Methyl	1.760	1.037	1.915	1.034	1.897	1.023	179.8	177.1	178.3
Ethyl	1.762	1.037	1.916	1.034	1.893	1.024	179.9	177.0	178.3
Propyl	1.762	1.037	1.915	1.034	1.891	1.024	179.8	177.1	178.3
Fluoro	1.743	1.040	1.915	1.032	1.932	1.022	177.9	176.8	178.0
Chloro	1.744	1.039	1.927	1.032	1.925	1.022	179.3	176.5	177.6
Bromo	1.740	1.039	1.925	1.032	1.923	1.022	179.6	176.6	177.8
Vinyl	1.759	1.037	1.921	1.033	1.909	1.022	179.6	176.7	177.9
Ethynyl	1.757	1.038	1.922	1.032	1.929	1.021	178.8	176.8	177.9
Propynyl	1.764	1.037	1.918	1.033	1.915	1.022	178.7	177.1	178.3
Phenyl	1.765	1.036	1.921	1.034	1.899	1.023	179.3	177.1	178.3
Pyridine	1.796	1.034	1.935	1.033	1.899	1.023	179.9	177.0	178.1
Pyrimidine	1.791	1.034	1.936	1.033	1.902	1.022	179.6	177.1	178.0
Quin	1.745	1.039	1.924	1.033	1.895	1.023	179.3	176.7	178.0
Cyano	1.730	1.042	1.933	1.030	1.967	1.019	179.2	175.9	177.0
Nitro	1.735	1.041	1.955	1.029	1.966	1.019	179.3	175.4	176.3
Methoxy	1.766	1.037	1.908	1.034	1.902	1.023	177.7	177.3	178.6

A summary of the binding energy for each analog with and without solvation is shown in Table 3.4. The table breaks down the contribution into electronic energy, enthalpy, entropic contribution, and free energy in the gas phase, and in aqueous solution using both PCM and COSMO solvation models. The binding enthalpy for the native

CG pair in the present study (-23.1 kcal/mol) is reasonably close to both the experimental value of Yanson an co-workers of -21.0 kcal/mol[233] as well as various theoretical calculations[234–236] including the value of -23.8 kcal/mol reported by Meng *et al.*[212]

Table 3.4: GC base pair binding electronic energy, enthalpy, entropic contribution, free energy, and free energy with solvation corrections. All values are give in kcal/mol.

Substituent	ΔE	ΔH	$-T\Delta S$	ΔG	ΔG_{aq}^{PCM}	ΔG_{aq}^{COSMO}
Cytosine	-24.8	-23.1	12.0	-11.2	10.8	11.0
Methyl	-25.3	-23.6	12.1	-11.5	10.5	10.4
Ethyl	-25.3	-23.7	11.8	-11.9	10.5	10.5
Propyl	-25.4	-23.7	11.8	-12.0	10.5	10.6
Fluoro	-24.5	-22.8	12.2	-10.6	11.3	11.4
Chloro	-24.2	-22.5	12.0	-10.4	11.3	11.4
Bromo	-24.1	-22.4	12.3	-10.1	11.6	11.8
Vinyl	-24.5	-22.9	11.7	-11.2	10.5	10.7
Ethynyl	-23.8	-22.2	12.0	-10.2	11.4	11.6
Propynyl	-24.3	-22.7	11.3	-11.4	10.6	10.7
Phenyl	-24.7	-23.0	11.9	-11.1	11.0	11.2
Pyridine	-23.4	-21.7	11.8	-9.9	11.2	11.3
Pyrimidine	-23.5	-21.8	11.9	-9.9	11.4	11.5
Quin	-25.1	-23.6	11.7	-11.9	10.2	10.5
Cyano	-22.8	-21.2	11.8	-9.4	11.7	11.7
Nitro	-21.8	-20.3	11.4	-8.9	11.7	11.7
Formyl	-22.0	-20.3	11.5	-8.9	11.5	11.5
Methoxy	-25.1	-23.4	12.2	-11.4	10.9	10.5
Oxymethyl	-23.8	-22.3	11.6	-10.7	10.5	10.2

The gas phase binding enthalpy values span a range of 3.5 kcal/mol. The trends are similar, but smaller in magnitude for the gas phase binding free energy. Solvation, which disfavors the nucleobase association, decreases this range to 1.6 kcal/mol. In the gas phase, the alkanes, propynyl, quin, and methoxy analogs all show increased binding for the GC base pair, while the halogens, ethynyl, pyridine, pyrimidine, formyl, oxymethyl, cyano, and nitro show decreased binding. This trend of electron withdrawing substituents destabilizing binding energy is consistent with work by Kawahara *et al.*[229] The phenyl and vinyl analogs show negligible change to free energy. Due to the overall

small changes in binding free energy, it is likely that the observed changes in melting temperature of DNA helices with these analogs[237] are only mildly affected by changes in base pair hydrogen bonding, and that other contributions, such as base stacking and solvation, are more significant.

3.3.3 Protonated CG Base Pair

Table 3.5: GC base pair hydrogen bonding geometry with N^2 protonated guanine. Bond lengths are given in Å. Angles are given in degrees.

	Bond Length						Angle		
Substituent	A	D	B	E	C	F	A/D	B/E	C/F
Cytosine	1.751	1.043	1.875	1.027	1.033	1.676	178.8	178.7	170.5
Methyl	1.750	1.042	1.879	1.028	1.029	1.689	179.4	178.6	170.5
Ethyl	1.753	1.042	1.881	1.028	1.028	1.692	179.7	178.6	170.5
Propyl	1.754	1.041	1.882	1.028	1.027	1.695	179.6	178.5	170.4
Fluoro	1.856	1.031	1.773	1.045	1.394	1.154	172.2	179.7	171.8
Chloro	1.849	1.031	1.780	1.046	1.387	1.157	173.8	179.6	171.8
Bromo	1.846	1.030	1.779	1.046	1.383	1.159	174.1	179.7	171.8
Vinyl	-	-	-	-	-	-	-	-	-
Ethynyl	1.862	1.030	1.773	1.047	1.375	1.164	173.1	180.0	171.6
Propynyl	1.762	1.041	1.876	1.028	1.029	1.688	178.1	178.8	170.6
Phenyl	1.764	1.040	1.883	1.028	1.028	1.692	178.9	178.7	170.3
Pyridine	-	-	-	-	-	-	-	-	-
Pyrimidine	1.794	1.037	1.893	1.028	1.030	1.681	179.2	179.1	170.3
Quin	1.733	1.044	1.882	1.029	1.026	1.696	179.1	178.5	169.8
Cyano	1.845	1.031	1.783	1.046	1.458	1.123	173.5	179.2	172.5
Nitro	1.847	1.031	1.804	1.044	1.466	1.119	175.0	178.9	172.8
Formyl	1.878	1.029	1.782	1.047	1.444	1.129	172.9	179.6	172.3
Methoxy	1.758	1.041	1.874	1.028	1.026	1.705	177.1	178.7	170.6
Oxymethyl	-	-	-	-	-	-	-	-	-

Table 3.5 gives the CG base pair geometry with the guanine N^2 protonated (Fig. 3.4). As discussed below, this protonation is a model reaction for electrophilic carcinogen attack. These structures are planar, first order transition states as discussed by

Dannenberg *et al.*[119] with an imaginary frequency attributed to inversion about the exocyclic, protonated N^2 amino group. Fully optimized structures of the protonated base pair are not planar in the gas phase making it difficult to compare to the planar bases within a DNA helix, therefore the analysis was restricted to the first order transition, planar structures. Values for vinyl and pyridine analog are not reported as the planar transition state structures were not found.

Compared to the CG base pairs, the protonated CG base pairs have much larger deviations from the native geometric values. The most interesting change observed is for the cytosine O^2 acceptor and guanine N^2 donor. For the alkyl, propynyl, aromatic, and methoxy analogs, the proton has completely transferred from the guanine amino group to the cytosine carbonyl. This proton transfer is very important as it may mimic proton movement during carcinogen nucleophilic attack at the guanine N^2.

PA, GPB, and relative pK_{a} values for the protonation of GC base pair at the guanine N^2 position are summarized in Table 3.6. Similar to previous the single cytosine results, the addition of solvation makes minimal changes to the analog trends and decreases the overall range of the deviation from the native GC base pair. The PA and GPB correlate well with the protonation of single cytosine O^2, see Table 3.2. This is a useful observation because the single cytosine base calculation is much less computationally intensive than the base pair and returned the same trend. Halogen, ethynyl, cyano, and nitro substituents lead to decreased PA and GPB, while alkanes, propynyl, aromatics, and methoxy increase PA and GPB. It should be noted that given the correlation between the single cytosine and base pair calculations, it is likely that vinyl and pyridine increase PA and GPB. Dannenberg and Tomasz reported the GPB for fluoro and methyl as -2.7 and 2.6 kcal/mol, respectively,[119] which compare well to our -3.6 and 2.5 kcal/mol.

3.4 Discussion

To determine how cytosine methylation affects reactivity of neighboring guanine, we consider likely ways the extra methyl group might influence each reaction step. This is particularly difficult given that various carcinogens may have significantly different mechanisms. In this study, we have analyzed the geometric and electronic properties of both the single cytosine and CG base pair with a variety of C5 substituents and

Table 3.6: Proton affinity ($\triangle H$), gas phase basicity ($\triangle G$) and relative pK_{a} (ΔpK_{a}) of N^2 guanine in the GC base pair. Proton affinity and gas phase basicity are given in kcal/mol. Relative pK_{a} values are calculated as the analogue pK_{a} minus the cytosine pK_{a}, see Eq. 3.3.

Substituent	($\triangle H$)	($\triangle G$)	Δp$K_{\mathrm{a}}^{\mathrm{PCM}}$	Δp$K_{\mathrm{a}}^{\mathrm{COSMO}}$
Cytosine	224.4	213.9	0.0	0.0
Methyl	226.7	216.4	0.4	0.3
Ethyl	227.6	217.2	0.7	0.7
Propyl	228.1	217.7	0.8	0.9
Fluoro	220.9	210.3	-1.6	-1.6
Chloro	221.1	210.4	-1.9	-1.9
Bromo	221.4	211.1	-1.7	-1.6
Vinyl	-	-	-	-
Ethynyl	222.1	211.5	-2.1	-2.0
Propynyl	226.2	216.7	-0.1	0.1
Phenyl	227.5	217.2	-0.1	0.1
Pyridine	-	-	-	-
Pyrimidine	228.1	217.8	-0.6	-0.6
Quin	227.2	216.7	-0.1	0.1
Cyano	215.8	205.2	-3.2	-3.2
Nitro	215.1	204.3	-3.7	-3.5
Formyl	218.8	208.1	-3.0	-3.0
Methoxy	228.3	217.8	0.9	0.5
Oxymethyl	-	-	-	-

protonation states. We can apply this knowledge to the possible reaction steps shown in Figure 3.1.

Geometric properties of cytosine and CG base pair analogs show minimal change compared to the native structures, likely indicating little impact on the local helical structure. Further, the C5 substituents protrude from the major groove, which is up to 12 Å wide for B DNA,[238] and therefore are unlikely to sterically distort the helix. On the other hand, the polarizability was increased by as much as a 100% upon C5 functionalization. While some polarizability of cytosine lies outside the base stacking region, this enhancement likely influences base stacking observed experimentally.[239]

Pre-covalent binding of reactive species to DNA likely precedes the chemical reaction, so substituent effects on the binding interaction will be a key factor in understanding the overall reaction. Cytosine analogs show a range of 13.0 to 24.1 $Å^3$ in polarizability and -0.192 to 0.168 e for total substituent charge. The range of the total substituent charges goes from electronically negative to electronically positive with a range of polarizabilities signifying each analogs creates a unique local electrostatic environment that may alter solvation and binding affinity. Further, Table 3.1 shows some of the substituents protrude as much as 4.3 Å farther out of the major groove than the native cytosine hydrogen, indicating this electronic environment is well presented to a possible reactive species. This indicates that each analog is likely creating subtly different pre-covalent binding sites for incoming carcinogens.

For some carcinogens, such as BPDE, intercalation is known to occur prior to chemical reaction.[240] The ability for a carcinogen to intercalate depends on the stacking ability of the bases and the ease of entry for the intercalator into the helix. If stacking ability is proportional to polarizability, as suggested previously,[241] then all of the analogs considered here (except for fluoro) should show increased base stacking. It is difficult to correlate this stacking increase with reactivity without knowing the relative Gibb's free energy between carcinogen intercalated (Figure 3.1 c) and outside the helix (Figure 3.1 b). If increased stacking lowers the intercalated complex energy more than the lone DNA duplex energy, then our analogs should increase intercalation. If the reactivity for a given carcinogen is proportional to the intercalation, one would also see increased reactivity.

While some C5 substituents are quite flexible (e.g. propyl) others are conformationally constrained (e.g. propynyl). The flexibility may thus influence the steric barrier to intercalation. The pyrimidine, pyridine, and phenyl analogs were chosen specifically to explore the intercalation enhancement along with changing steric barrier. If enhanced polarizability correlates with increased pre-covalent retention times and reactivity and steric barrier to intercalation plays no role, then the pyridine, pyrimidine, and phenyl, quin analogues will show significant reactivity increases. If instead the relative reactivity is shown to follow quin > pyrimidine > pyridine > phenyl, then steric obstruction to intercalation is likely occurring.

Experimental results for the reactivity changes of mytomycin C using the fluoro and methyl analogs been attributed to a transmitted electronic effect from the substituent through the hydrogen bonds to the guanine N^2 amino group.[199] In essence, the substituent donates or withdraws its nucleophilicity through the cytosine O^2 to the guanine N^2 resulting in an increase or decrease in reactivity toward electrophiles. Previous calculations[119] used protonation of the guanine N^2 in the CG base pair as an indicator of the strength of this effect. These results showed methyl favoring protonation by 2.6 kcal/mol and fluoro disfavoring the reaction by 2.7 kcal/mol compared to the native cytosine free energy. We see similar results, 2.5 and 3.6 kcal/mol for methyl and fluoro, respectively.

If this electronic effect plays the dominant role in the reaction, then by inference, the present results would suggest the following order of analog reactivity toward electrophiles:

methoxy, pyridine, pyrimidine, propyl, phenyl, ethyl >
quin, propynyl,oxymethyl, methyl, vinyl >
native >
ethynyl, bromo > chloro, fluoro > formyl >cyano, nitro

(see Table 3.6). We have grouped substituents with similar values based on the errors known for the QCRNA protocol (the standard deviation for relative GPB values is ∼1 kcal/mol).[139,140] The placement of vinyl, oxymethyl, and pyridine was estimated from the single cytosine gas phase basicities, which generally mirrors the base pair GPB. This trend is also seen the work by Dannenberg *et al.*[119]

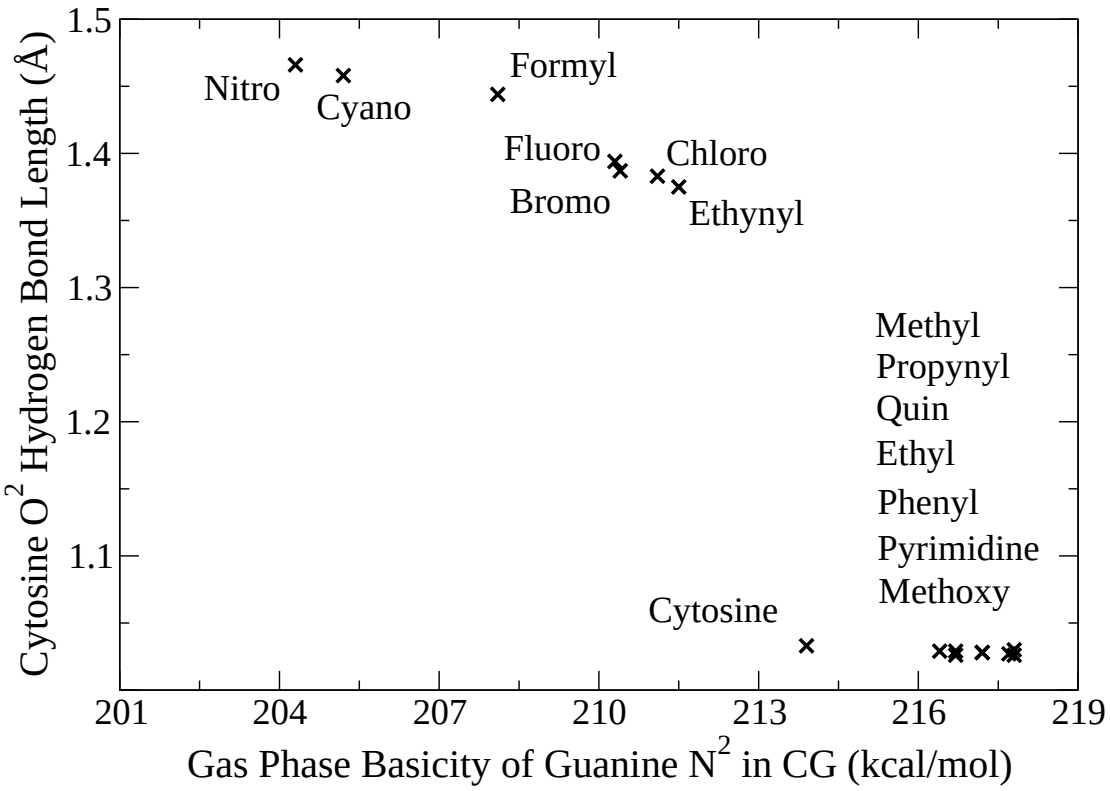

Figure 3.5: Gas phase basicity and hydrogen bond distance between cytosine O^2 and guanine N^2 after protonation

The electronic effect is also consistent with the changes in hydrogen bond distances for the protonated GC. Figure 3.5 plots the base pair GPB versus the hydrogen bond distance between the cytosine O^2 and the proton within the hydrogen bond to the guanine N^2. Results indicate that substituents that donate nucleophilicity enhance this proton transfer in the protonated base pair, while electron withdrawing substituents will disfavor proton transfer. In summary, if the electronic effect is a significant factor in the chemical steps of the reaction, analogs that can donate nucleophilicity through the hydrogen bonds should increase the GPB, facilitate proton transfer to the cytosine carbonyl, and increase reactivity at the guanine amine.

In order to fully probe chemical mechanism with theoretical methods, for even a particular carcinogen, a multi-faceted approach is required that includes molecular simulation, quantum chemical calculations, and correlation with experimental data. Nonetheless, a first step toward providing insight into reactivity of methylated CG base pairs as dictated by experimental chemical modification of cytosine, is to characterize substitution effects on the electronic structure of the individual base and base pair. The data provided herein establishes a baseline characterization across a broad range of chemical modifications that may help to guide and interpret the results of future experiments.

3.5 Conclusion

Cytosine methylation at C5 is an important epigenetic base modification that influences gene expression[186] and is correlated to known lung cancer mutational hot spots in the *p53* tumor suppressor gene.[188] While it is widely known that this relatively small chemical modification significantly changes reactivity of the base paired guanine toward electrophiles, the actual mechanism is both unknown and likely composed of many steps. We have proposed a variety of C5 cytosine substitutions chosen specifically to elucidate the relative importance of likely reaction steps.

In this work, we used density functional theory to calculate various cytosine analog properties of both the single cytosine bases and in the CG base pairs. These properties include hydrogen bonding structure, atomic charges, polarizability, proton affinity, gas phase basicity, and pK_a. Our results show that the cytosine analogs present a distinct

electronic environment to an incoming carcinogen, which will likely influence the pre-covalent major groove binding. Further, previous work suggested that reactivity changes are based on an electronic effect that is transferred from the substituent through the cytosine carbonyl to the guanine amino group. We have quantified this effect for our analogs such that future experimental work can conclude the importance of this electronic effect. As whole, these results provide a guide to future interpretation into the relative importance of the reaction steps required for carcinogen and drug reactivity at methylated CG steps.

Chapter 4

Exocycle lesions of adenine

This chapter contains excerpts from "Exocyclic Adducts Deoxyadenosine of 1,2,3,4-Diepoxybutane: Synthesis, Structural Elucidation, and Mechanistic Studies", which was submitted for review to the Journal of the American Chemical Society. This is a collaborative work between the Tretyakova Lab (Uthpala Seneviratne, Sergey Antsypovich, Danae Quirk Dorrm, Rebecca Guza, Melissa Goggin, Carrie Thompon and Professor Natalia Tretyakova) and the York Lab (Adam Moser and Professor Darrin M. York). The following focuses on the computational contribution to the work and does not include all the experimental results.

4.1 Introduction

Exocyclic nucleobase adducts are among the most important types of DNA damage because of their ability to exert significant biological effects.[242–244] These bifunctional lesions are characterized by considerable changes of the molecular size/shape and hydrogen bonding characteristics of the parent nucleobase, leading to mispairing during DNA synthesis.[245–250] For example, N^6-etheno-deoxyadenosine adducts induced by vinyl chloride preferentially adopt the syn conformation about the glycosidic bond, forming a Hoogsteen base pair with guanine or cytosine instead of the normal adenine partner, thymine[245,250]

One prominent bis-electrophile capable of inducing exocyclic nucleobase lesions is 1,2,3,4-diepoxybutane (DEB), the ultimate carcinogenic metabolite of 1,3-butadiene.[242]

1,3-Butadiene is a known animal and human carcinogen found in automobile exhaust and in cigarette smoke.[193,251] The adverse biological effects of DEB have been attributed to its ability to cross-link cellular biomolecules. Initial alkylation of adenine and guanine bases in DNA by DEB produces 2-hydroxy-3,4-epoxybut-1-yl (HEB) lesions, which contain an inherently reactive oxirane group and can alkylate neighboring nucleobases within the DNA duplex to form DNA-DNA cross links.[252,253] Alternatively, the 3,4-epoxy ring can be subject to nucleophilic attack by another site within the same DNA nucleobase, giving rise to fused ring structures.[254–256]

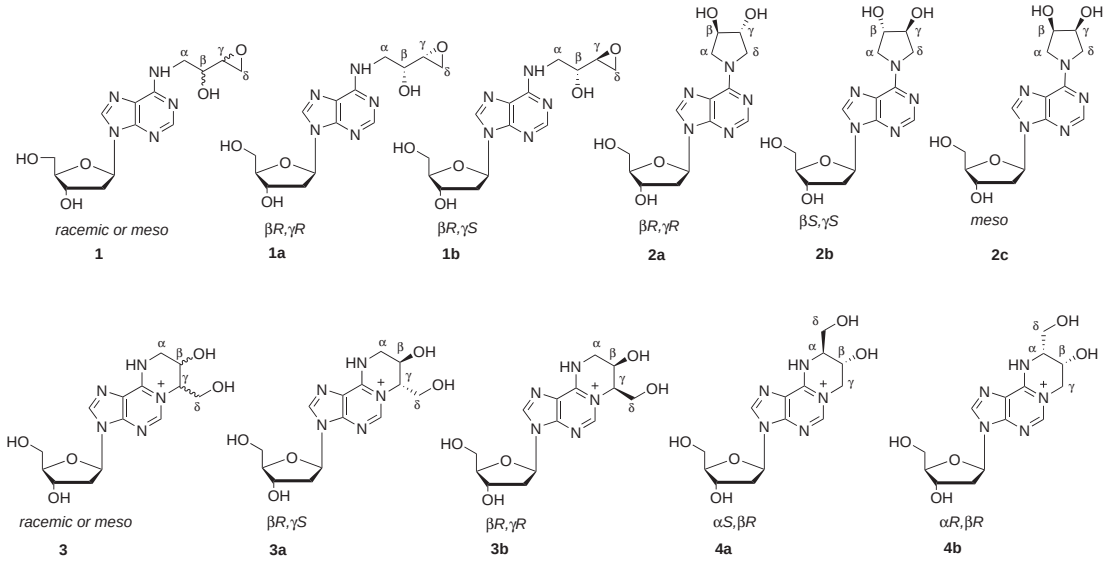

Figure 4.1: dA lesions from 1,2,3,4-diepoxybutane attack.

The documented ability of DEB to induce large numbers of A → T transversion mutations[257] has led us to hypothesize that it forms strongly mispairing exocyclic lesions at adenine nucleobases within DNA. This hypothesis was supported by our previous work with synthetic DNA oligonucleotides containing site specific N^6-(2-hydroxy-3,4-epoxybut-1-yl)deoxyadenosine adducts (N^6-HEB-dA, 1 in Figure 4.1)[254] If left in an aqueous solution at room temperature (pH 7.2), N^6-HEB-dA underwent spontaneous cyclization to form previously unidentified DEB-dA lesions.[254] Another isomer of the

exocyclic DEB-dA species was formed as a side product during the synthesis of N^6-HEB-dA by reacting 6-chloropurine deoxyriboside with 1-amino-2-hydroxy-3,4-epoxybutane under basic, anhydrous conditions.[254] In the present work, a combination of UV and NMR spectroscopy, tandem mass spectrometry, independent synthesis, DFT calculations, and kinetic analysis was employed to identify the chemical structures of these novel DEB-DNA lesions and to establish the mechanism of their formation.

4.2 Computational Methods

Density-functional electronic structure calculations were performed according to the protocol used in the QCRNA database.[88] The calculations included geometry optimization, vibrational frequency analysis, and solvation energy corrections. Kohn-Sham density-functional calculations were performed using the hybrid exchange functional of Becke[62,63] and the Lee, Yang, and Parr correlation functional[64] (B3LYP) as implemented in the Gaussian03 suite of programs.[122] Solvation corrections were calculated using the COSMO,[82] and PCM[225] solvation models. Relaxed potential energy surface scans around the C-N torsion angle were performed at the B3LYP/6-31+G(d,p) level with the torsion constrain set in 15 degree intervals, with additional unconstrained (fully optimized) points at the minima.

4.3 Results and Discussion

4.3.1 Structural determination of N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine

The independent synthesis of adducts **2a**, **2b**, and **2c** in Figure 4.1 were completed and HPLC-UV-ESI+MS and periodate oxidation was used to verify the products, in particular the 5 membered ring structure and existence of the diol. To further verify the structure proton and carbon 1D and 2D NMR was performed. The initial NMR of **2a** at r.t. is shown in Figure 4.6 (top frame).

The asymmetric nature of the proton NMR revealed that the 5-membered lesion might not be fixed conformationally with respect to the adenine aromatic system. To

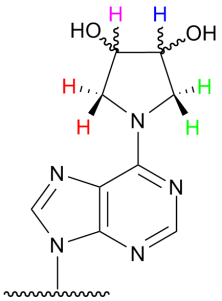

Figure 4.2: N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine. Proton α (red), β (pink), γ (blue), and δ (green).

investigate this quantum calculations were first carried out to determine what conformations are available to the 5-membered lesion and then the rotational barrier of the 5-membered ring with respect to the adenine.

Optimization and analysis of 2a, 2b, and 2c bases from Figure 4.1 were performed. Here we focus on the 2a as the trends are similar for each steriochemistry. Initial optimization of the steriochemical 2a-base (βR,γR) revealed two possible sugar pucker states for the 5-membered ring. Shown in Figure 4.3, these two conformations distinguished most easily by the orientation of the β and γ hydroxyl groups, which can either both be in the axial or equatorial positions. The axial conformation was found to be $\sim$1 kcal/mol lower in free energy than the equatorial. Addition of solvation corrections with both the COSMO and PCM implicit solvation did not change the relative energy. This indicates there will be significant amount of each pucker if the transition barrier is low. The transition state for this pucker transition was determined and a barrier of $\sim$4 kcal/mol in free energy was found. Again solvation did not change this value significantly.

Molecular orbitals for the axial, equatorial, and pucker transition state structures of 2a-base were visualized. All structures showed at least 7 orbitals with significant density above and below the C-N σ bond that connects the 5-membered ring and the adenine. Figure 4.4 shows presents two of these orbitals to visualize this point: a top view and side view of HOMO-2 and HOMO-20 of compound 2a-base in an equatorial-like conformation. The electron density in the π region of the C-N bond indicate the nitrogen

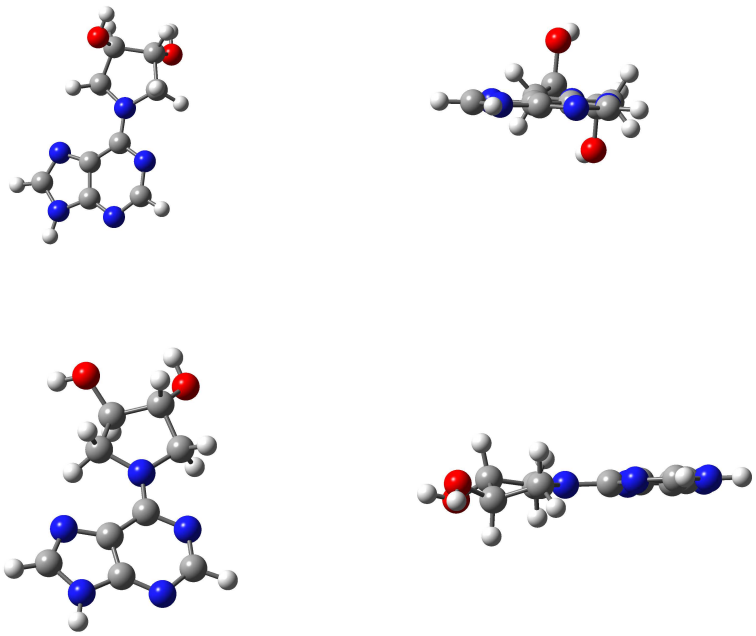

Figure 4.3: Axial (top) and equatorial (bottom) conformations of R,R N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine with top and side view.

is exhibiting significant sp^2 like character even though it is formally sp^3. Natural bond order analysis gives a bond order of 1.2 for the C-N bond, further supporting the sp^2 nature of the bond. To investigate the affect of this on the relative ring rotations a relaxed potential energy scan of this motion was performed.

Figure 4.5 shows the relaxed potential energy surface scan of the C-N dihedral that controls the relative rotation of the five membered ring and the adenine. Based on the lewis structure of 2a, it might be expected the surface would be symmetric about 180.0 degrees, but the puckering of the 5-membered lesion creates an asymmetric surface because of the orientation of the β and γ hydroxyl groups. The estimated barrier is $\sim$14.5 kcal/mol for the axial and $\sim$17.0 kcal/mol for the equatorial. This rotational barrier, while high enough to inhibit rotation on the NMR timescale at r.t., was small

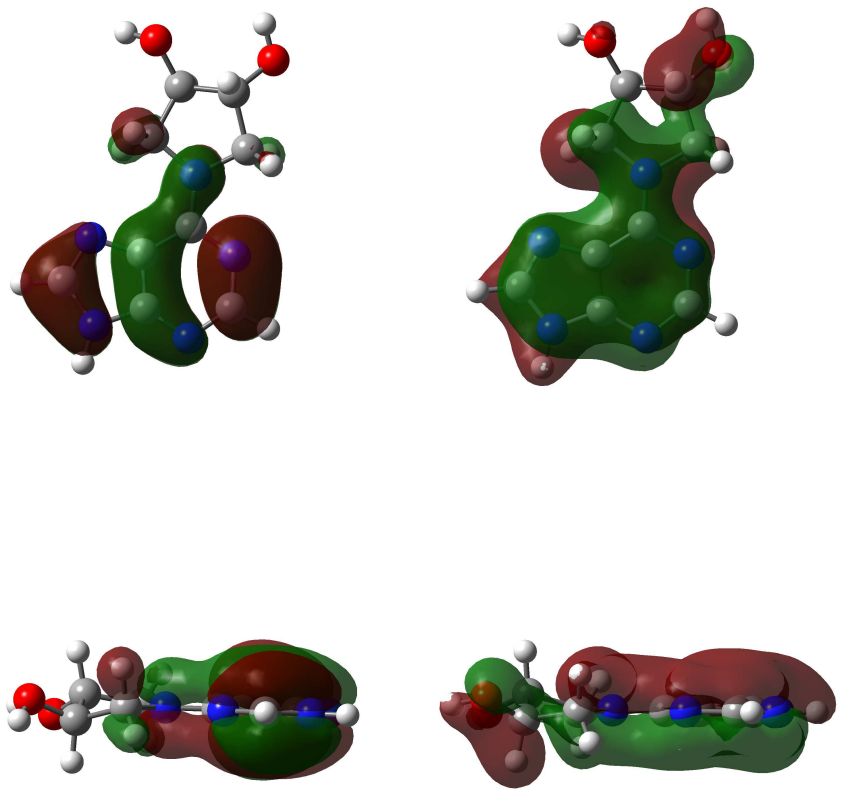

Figure 4.4: Molecular orbitals of N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine in the equatorial conformation. HOMO-2 top view (top, left) and side view (bottom, left). HOMO-20 top view (top, right) and side view (bottom, right).

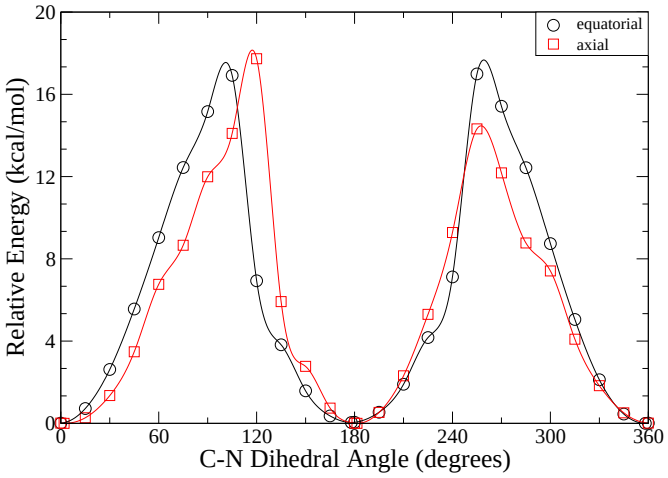

Figure 4.5: Relaxed potential energy scan of the rotational barrier of R,R-N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine in the equatorial (circle) and axial (square) conformation around the C-N dihedral between the five and six membered rings. Continuous lines are spline fits to the data.

enough to predict significant rotation at slightly elevated temperatures. This lead to temperature dependant NMR being preformed, the results are shown in Figure 4.6.

At low temperature, none of the protons are equivalent which is consistent with the rotational barrier calculated above and the low barrier for pucker transition. As the temperature increases the α and δ protons merge as well as the β and γ protons, which is consistent with the computational prediction that as the C-N dihedral is able to rotate, these protons should become conformationally indistinguishable and therefore feel the same NMR deshielding. These results are similar for 2b (S,S stereochemistry) in Figure 4.1. 2c, which is a meso mixture of R,S and S,R stereochemistries for the β and γ positions. At high temperatures all the conformations are accessible from any by either a pucker transition or ring rotation followed by a pucker transition. This is illustrated in Figure 4.7.

4.4 Conclusion

This is the first report of the formation of potentially promutagenic exocyclic DEB-dA lesions. The structural identities and optical configurations of the novel nucleosides were confirmed by independent synthesis, NMR, MS, and DFT calculations. In this work, computational chemistry techniques, in particular density functional calculations, were used to help explain both NMR spectra and the product conformations.

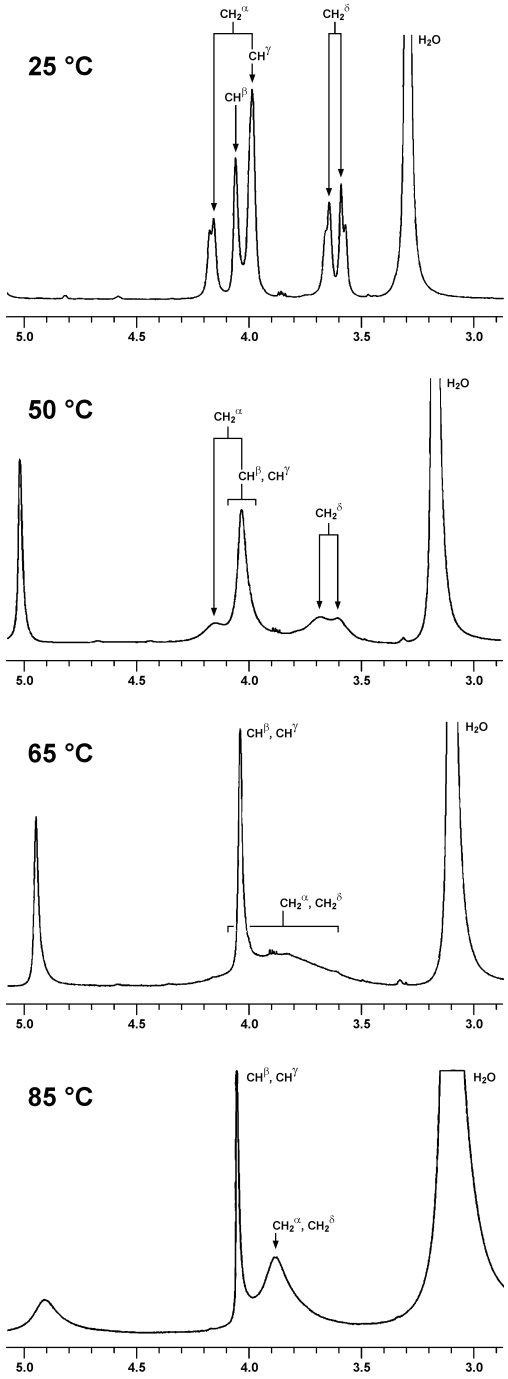

Figure 4.6: Temperature dependent proton NMR of R,R -N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine at 25 °C, 50°C, 65°C, and 80°C with labeled peaks.

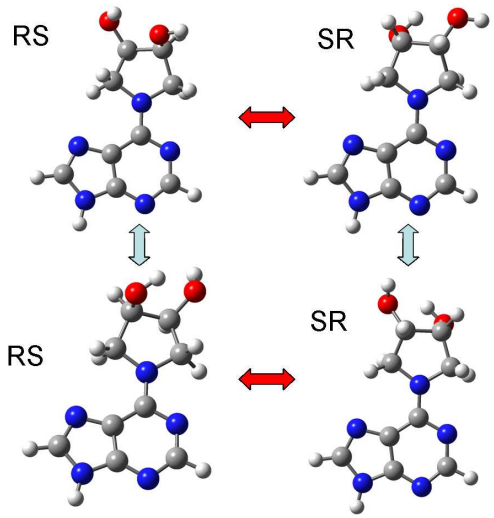

Figure 4.7: Conformational changes between the four stereochemistries of meso N^6,N^6-(2,3-dihydroxybutan-1,4-diyl)-2'-deoxyadenosine (2c). Red arrows indicate ring rotation. Blue arrows indicate pucker transitions.

Chapter 5

Parameterization

While small molecule quantum studies provide excellent benchmarks and insight into biomolecules, many problems of biological interest are of systems much larger than is feasible using quantum mechanics. In this case, different models, like the molecular mechanics models described in Chapter 1, can provide the scaling and computational efficiency to attack these questions. Experimental data can be used to parameterize, but often the data does not exist or cannot be sufficiently disconnected from experimental factors to be used. In this case, quantum mechanical calculations can provide the necessary rigor to develop molecular mechanics models. This approach has been used by many of the most successful molecular mechanics force fields.[258–262] Here two applications are detailed: CHARMM force field parameters for reactive intermediates of ribozymes and phosphate solvation benchmark data.

5.1 CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes

The section contains published work in part from "CHARMM Force Field Parameters for Simulation of Reactive Intermediates in Native and Thio-Substituted Ribozymes" E. Mayaan, A. Moser, A. D. MacKerell Jr., D. M. York, J. Comput. Chem. **28**, 495 (2007).[263]

5.1.1 Background

Over the last several decades a wealth of data has accumulated that demonstrates the central role RNA catalysis plays in many biological processes. Starting in the late 1970's, it was shown that RNA could catalyze fairly complex biological reactions in ribonuclease P[264,265] and Tetrahymena[266,267] with an efficiency that rivaled many protein enzymes. These discoveries sparked a wave of interest in the scientific community focused on unraveling the details of how RNA enzymes (ribozymes) function. An understanding of the catalytic mechanisms of ribozymes, and their relation to sequence and tertiary structure, is opening up a variety of new frontiers. In biomedical technology, gene expression inhibitors that target viral and genetic diseases[13] such as HIV[6] and cancer[9] are being developed, and new biotechnologies such as RNA chips[39] and allosteric molecular switches in nanodevices[24] are being explored.

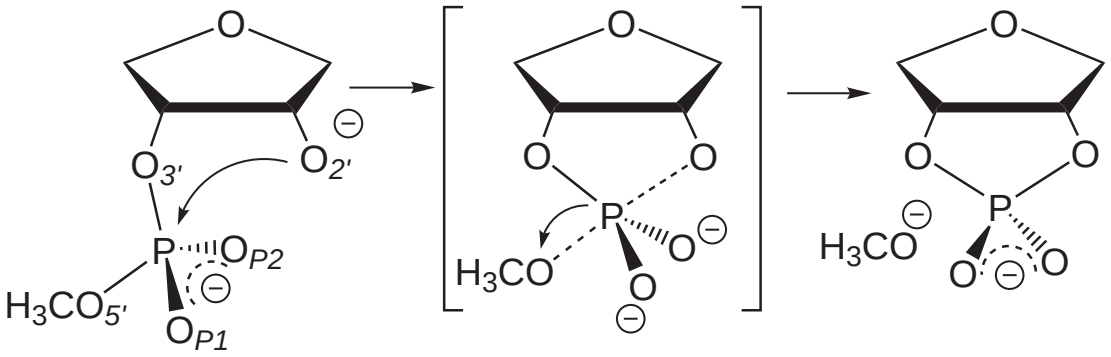

Figure 5.1: Model RNA transesterification reaction.

A common biological reaction catalyzed by several prototype ribozyme systems, such as the hammerhead,[268,269] hairpin,[45,270] and hepatitis delta virus[271,272] ribozymes, involves cleavage of a phosphate group through a transesterification reaction[273,274] (Figure 5.1). In this reaction, the 2'OH of the ribose sugar ring becomes activated via deprotonation and makes an in-line attack to the adjacent 3'-phosphate along the phosphodiester backbone. The attack produces a trigonal bipyramidal phosphorane intermediate/transition state that is accompanied by an inversion of configuration about the

phosphorus center as the exocyclic P-O5' bond is cleaved. The product of the transesterification is a 5'-OH terminus and a 2',3'-cyclic phosphate. Additional information about this mechanism has been obtained via kinetic isotope studies[275–277] and chemical modifications such as thio-substitution[278–280] at the scissile phosphate, although the mechanistic interpretation of these experimental studies remains a topic of discussion and some debate. Time-resolved x-ray crystallography[268,281,282] has become a powerful tool to elucidate structural information at different stages along the catalytic reaction coordinate that provides valuable insight into ribozyme activity. However, this area is considerably challenging due to the difficulty of trapping a reactive intermediate and obtaining quality crystals, as well as uncertainties due to the nature of effects that arise from the crystallization conditions. For these reasons, there is considerable interest in the development of theoretical methods that can aid in the refinement and interpretation of existing experimental data, and provide structural insight into systems where such data is not yet available.

Molecular simulation, along with experimental structural data, provides an avenue for the characterization of ribozyme dynamics in solution and refinement of key mechanistic details. Molecular simulation force fields for nucleic acids continue to improve[283–288] and a variety of simulations involving ribozymes have been carried out in recent years.[289–297] In order to study the structure and dynamics of different catalytic states along the reaction path of a ribozyme, however, reliable empirical force fields parameters must be developed for the transition states and reactive intermediates of these reactions. Furthermore, in order to use molecular simulation to aid in the interpretation of experimentally measured thio effects, parameters for thio-substituted phosphate and phosphorane models and their interactions with metal ions are required. In the present work, new force field parameters for residues important to the study of RNA catalysis are derived from density-functional calculations to be consistent with the CHARMM27[261] all-atom empirical force field. These parameters will allow molecular dynamics (MD) simulations of ribozymes in reactive states to be performed to study the structure and dynamics that lead to catalysis.

5.1.2 Methods

The potential energy function used for the CHARMM27 empirical force field for nucleic acids,[262, 298] and for the new modified RNA residues of the present work, has the general form:[299]

$$\begin{aligned} U(\mathbf{r}_1, \mathbf{r}_2 \cdots \mathbf{r}_N) &= \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \\ &+ \sum_{UB} K_{UB}(S - S_0)^2 + \sum_{impropers} K_\varphi(\varphi - \varphi_0)^2 \\ &+ \sum_{dihedrals} K_\chi \left[1 + cos(n\chi - \chi_0)\right] \\ &+ \sum_{i,j<i} \epsilon_{ij} \left[\left(\frac{R_{0,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{0,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \end{aligned} \tag{5.1}$$

The first four summations are quadratic terms that give rise to energy penalties for geometrical deviations about equilibrium coordinate values. The variables b, θ, S, and φ are the bond length, bond angle, Urey-Bradley 1,3-distance, and improper torsion angle coordinates, respectively, while b_0, θ_0, S_0, φ_0 and K_b, K_θ, K_S, K_φ are the corresponding force field parameters for the equilibrium geometries and force constants, respectively. The fifth summation is a trigonometric term that adjusts the periodic dihedral angle rotational barriers. The coordinate χ is the dihedral coordinate, n determines the periodicity, χ_0 is a phase factor, and K_χ is the amplitude force constant. The terms involving a sum over atom pairs $i, j < i$ (neglecting non-bonded exclusions) are the non-bonded van der Waals/Lennard-Jones (L-J) and electrostatic terms. The parameters ϵ_{ij} and $R_{0,ij}$ are the van der Waals well depth and minimum distance between the ij atom pair, respectively, and are by default calculated via the Lorentz-Berthelot combining rules[300] from the corresponding 1-body parameters as $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and $R_{0,ij} = R_{0,i} + R_{0,j}$, respectively, although this default can be explicitly overridden using the NBFIX (non-bond fix) option if fine tuning of specific pairwise interactions is desired. The last summation is the electrostatic energy determined between atomic partial charges q_i, which normally are calculated with a unit dielectric constant, $\epsilon = 1$, as in the present work.

Parameterization of the CHARMM27 force field was based on *ab initio* and experimental data for small molecules,[262] as well as macromolecular simulation data of DNA

and RNA[301] in order to more accurately capture experimental condensed phase properties. This was an important improvement to the CHARMM22 force field[302] that was parameterized more heavily on the basis of small molecule target data alone. While this previous approach succeeded in capturing the target data of the selected small compounds, later simulations of duplex DNA in solution showed some disagreement with experiment in regard to the relative conformational stability of A-DNA versus B-DNA.[303,304]

The CHARMM27 force field was developed using a self-consistent, step-wise strategy.[262,288] Partial atomic charges for the CHARMM27 force field were initially obtained from a Mulliken population analysis of the HF/6-31G(d) wave function. These were then adjusted to better reproduce scaled HF/6-31G(d) TIP3P[70] water interaction energies, experimental dipole moments, and heats of sublimation where available. Equilibrium geometry parameters for the selected small molecules were initially optimized to reproduce the experimental geometries of microwave, electron diffraction and/or x-ray crystal survey data where possible. Equilibrium geometries and force constants were optimized iteratively until satisfactory fitting to the target data was achieved. Once this optimization was complete, iterative adjustment of the charges and L-J parameters was coupled with the internal parameters until overall convergence was reached. A survey of RNA and DNA crystal structures from the Nucleic Acids Database (NDB)[305] was taken so that fitting to target macromolecular experimental properties, such as sugar puckering phase and dihedral angle distributions, could be made through crystal MD simulation. During this stage, in the original CHARMM27 all-atom empirical force field parameterization for nucleic acids,[262] dihedral angle parameters were adjusted to lower or raise energy barriers such that target condensed phase properties were better reproduced. While this sometimes sacrificed the quality of fitting to small molecule ab initio data, it accomplished the goal of more accurately reproducing experimental condensed phase properties. This macromolecular fitting was coupled with the internal parameter optimization until satisfactory convergence was achieved.

Due to the lack of high-resolution experimental data for RNA reactive intermediates and chemically modified nucleic acids, the parameterization of the present study is based solely on ab initio data, the optimized structures for which are shown in Figures 5.4, 5.5, 5.6, and 5.7 at the end of this section. Parameter optimization for

modified CHARMM27 RNA, Mg^{2+}, and OH^- residues was based on density functional theory (DFT) calculations for training sets of small molecules that represented the desired target systems. Density-functional calculations were performed using the Becke three-parameter hybrid functional combined with the Lee-Yang-Parr exchange-correlation functional (B3LYP)[63,64] with the 6–31++G(d,p) basis set for geometry and frequency calculations followed by single-point electronic structure refinement with the 6–311++G(3df,2p) basis set in a manner analogous to recent studies of biological phosphates.[97,128–130,306] The B3LYP model employed here neglects proper treatment of long-range dispersion interactions; however, these weak dispersion interactions are small in comparison with the highly polar (e.g., hydrogen bonded) and ionic interactions investigated here, for which B3LYP has been demonstrated previously to be generally reliable. All DFT calculations were performed using the GAUSSIAN03[122] package. Partial atomic charges were based on the CHelpG[307,308] method (CHarges from ELectrostatic Potentials using a Grid) which fits atomic centered point charges to the molecular electrostatic potential. The CHelpG charges were also constrained to reproduce molecular dipole moments. Similar approaches to force field charge determination have been applied and validated previously.[307,309] These calculations were performed with the 6–311++G(3df,2p) basis set, and the CHelpG charges were subsequently modified to be consistent with the original force field by taking into account only the charge *differences* from standard CHARMM residues, and making further adjustments for polar and non-polar hydrogens and integer charged groups in accord with the original parameterization procedure.[262,288] The CHelpG charges used in the present work to obtain the charge differences with respect to similar standard CHARMM residues are different than the Mulliken charges that were used as a starting point for optimization in the original CHARMM27 force field, and in some cases these charge models differ significantly. As in the original CHARMM27 force field, transferring charges from the small model compounds to the nucleic acid fragments was accomplished by adding the charge of the removed hydrogen atom to the heavy atom from which it was deleted. Once charges were obtained, Lennard-Jones (L-J) parameters were determined by fitting CHARMM residue-TIP3P water interaction energies and distances to the density-functional target values. This was accomplished through scaling of the interaction energies and shifting of the binding distances such that the DFT values matched those of the original

CHARMM27 force field for unsubstituted dimethyl phosphate (DMP^-). Although the original CHARMM27 force field parameterization fixed waters to the TIP3P geometry in ab initio calculations, it was found in this work that the relative differences in geometries predicted by the DFT level of theory used compared better with the HF/6-31G(d) used in the original force field when the waters remained unconstrained. Internal parameters were fit to the density-functional geometries and energies for the model compounds.

The CHelpG charges used in the present work to obtain the charge differences with respect to similar standard CHARMM residues are different than the Mulliken charges that were used as a starting point for optimization in the original CHARMM27 force field. In general, the CHelpG charges are larger in magnitude, and due to the constraints, preserve the molecular dipole moments. The Mulliken charges on the other hand are considerably basis set dependent, particularly when diffuse functions are used, and generally smaller in magnitude. It should be emphasized that the Mulliken charges used as a starting point for parameter optimization with CHARMM27 were often significantly altered upon optimization to obtain intermolecular interactions with TIP3P water molecules, the final charges in many cases being closer to the CHelpG charges (e.g., for the non-bridging phosphate oxygens).

RNA generally has more accessible "folded" conformations than DNA[310] and often conformational deformation in the phosphate backbone is requisite for catalysis.[311] The CHARMM27 force field contains dihedral parameters for unsubstituted RNA based on *ab initio* gas-phase calculations of a variety of test compounds and adjusted with respect to A and B form DNA and RNA within the Nucleic Acid Database (NDB).[305] These adjustments included the lowering of torsional energy barriers in certain regions after the observation that direct parameterization to the *ab initio* torsion profiles resulted in simulated dihedral distributions that were inconsistent (e.g., too rigid) compared with those from a large NDB survey of RNA and DNA.[262] In the case of non-bridging thio-substitutions, no parameters exist. Therefore, attention was given to the C-O-P-O (i.e., α and ζ in Figure 5.2) dihedral parameters because of their role along the phosphate backbone and the possibility of significant change upon thio-substitution. As in CHARMM27, dimethyl phosphate was used as the model compound for native

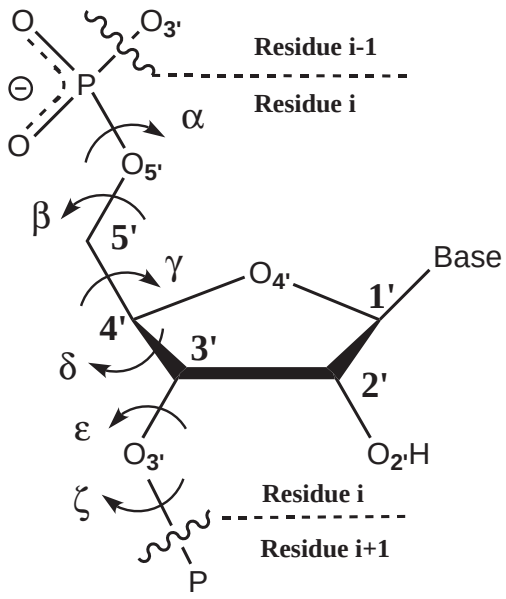

Figure 5.2: Phosphate backbone of RNA with torsions labeled.

RNA as well as singly and doubly thio-substituted dimethyl phosphate to determine the effects of the non-bridging sulfur(s) on this torsion. The dimethyl phosphate and thio-substituted torsional potential energy surfaces were generated using B3LYP/6-311++G(3df,2p)//B3LYP/6-31++G(d,p) with all degrees of freedom relaxed except for the dihedral of interest and the symmetric ζ dihedral, which is fixed in the *trans* conformation to induce symmetry and facilitate parameterization (Fig. 5.3).

As in the original force field parameterization,[262] a self-consistent step-wise optimization approach was taken that involved the iterative adjustments of L-J and internal parameters (not including torsion parameters) until convergence of the fitting function was obtained. Initial force constant values were taken from harmonic fitting to potential energy surface scans of bonds and angles with distortions of 0.2 Å and 2.0^o respectively. Starting geometries for the new CHARMM* model compounds were taken directly from the density-functional results and read into CHARMM to perform an Adopted Basis Newton-Raphson (ABNR) minimization[312] until a gradient of $<10^{-6}$ kcal/mol·Å was reached. The results of the CHARMM minimization with the current parameters were

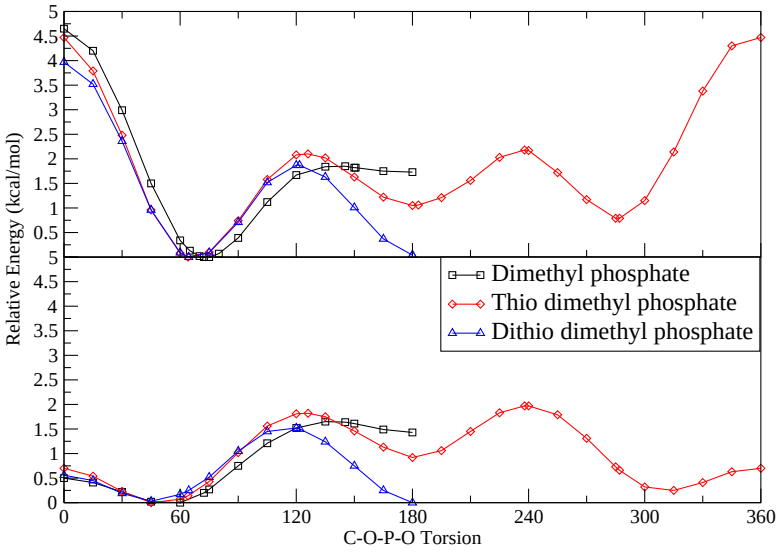

Figure 5.3: *Ab initio* (top frame) and CHARMM (bottom frame) torsional potential energy surface for the C-O-P-O dihedral of dimethyl phosphate (OO), non-bridging thiosubstituted dimethyl phosphate (SO), and non-bridging dithiosubstituted dimethyl phosphate (SS).

used to construct a χ^2 fitting function. In the present work, this function took the form:

$$\chi^2(\lambda) = \sum_{k=1}^{bonds}\left(\frac{b_k^* - b_k^{DFT}}{\sigma_{b,k}^2}\right)^2 + \sum_{k=1}^{angles}\left(\frac{\theta_k^* - \theta_k^{DFT}}{\sigma_{\theta,k}^2}\right)^2 + \sum_{k=1}^{torsions}\left(\frac{\chi_k^* - \chi_k^{DFT}}{\sigma_{\chi,k}^2}\right)^2 + \sum_{k=1}^{bind}\left(\frac{\Delta E_k^* - \Delta E_k^{DFT}}{\sigma_{E,k}^2}\right)^2 \quad (5.2)$$

Where the b_k^*, θ_k^*, χ_k^* and ΔE_k^* are the CHARMM energy-minimized bond lengths (Å), bond angles (degrees), torsion angles (degrees) and interaction energies (kcal/mol), respectively, and b_k^{DFT}, θ_k^{DFT}, χ_k^{DFT} and ΔE_k^{DFT} are the corresponding DFT reference values. The σ parameters, have the same units as their associated geometrical or energetic quantities (indicated by superscript), the inverse squares for which serve as weights in the unitless $\chi^2(\lambda)$ merit function. A non-linear optimization procedure is then used to minimize the $\chi^2(\lambda)$ merit function with respect to the CHARMM* parameters (this work) denoted by the vector λ. The σ parameters (and hence the weights) were adjusted empirically over the course of a stepwise minimization procedure in order to drive the

parameter search in the most realistic and robust region of parameter space (see Supporting Information for for more details). In most cases, the last stepwise minimization procedure used to arrive at the final set of parameters was a quadratically convergent direction set minimization method[313] that does not require explicit gradient information. Sets of conjugate directions were generated with the algorithm due to Powell in a series of successive line minimizations achieved by parabolic interpolation. The convergence criterion was 10^{-4} on the relative change in the function value with respect to its minimum value after a series of line minimizations. While this method works very well for finding the local minima of a multi-variable function, it is limited in its ability to find the absolute minimum of such a function. For this reason, the minimization procedure was carried out at a large number of starting points for each variable until the best overall minima could be found. Once a minimum was reached the torsion angle parameters were adjusted empirically to obtain a reasonable balance between the trends in the relative ab initio curves (native versus sulfur substituted) while maintaining consistency with the standard CHARMM27 torsion profile which does not correspond directly to an *ab initio* curve. Course grained adjustment of the parameters (0.05 kcal) was seen to be sufficient to provide an acceptable balance. The torsion parameters were adjusted iteratively with the parameters of the non-linear optimization.

5.1.3 Results and Discussion

Magnesium Complexes

RNA enzymes show diversity in the number as well as the role of metal ions they require. Metal ions can have both structural and chemical roles in catalysis. Possible (chemical) catalytic roles for metal ions include acting as:[269,274]

- General acid-base catalyst by accepting/donating protons from coordinated water molecules to the nucleophilic (2') and leaving group (5') positions
- Lewis acid catalyst by direct coordination to either the nucleophilic (2') or leaving group (5') positions

- Electrophilic catalyst by coordinating to the non-bridging phosphoryl oxygens and assisting in formation of a phosphorane-like transition state or intermediate

One way for a metal ion to act as a general base is to abstract a proton from a coordinated water to form a metal hydroxide. This OH^- has been postulated to play the role of abstracting a proton from a ribose 2'-OH,[314] although the p$K_a$difference between a phosphate coordinated Mg^{2+} and a ribose 2'OH remains a topic of current interest.[108] In one postulated two metal ion mechanism for the hammerhead ribozyme (HR), it has been proposed that a metal hydroxide exists as a μ-bridging OH^- between two magnesium ions.[296] Other studies suggest a single metal ion mechanism with an outer-sphere coordination of the Mg^{2+} ion to the 2'OH[315] and still others suggests a non-bridging two metal ion mechanism with inner-sphere binding of Mg^{2+}.[316] Clearly, MD and QM/MM simulation could help resolve these mechanisms and the catalytic role played by Mg^{2+} ions.

The force field parameters for Mg^{2+} and OH^- are important to obtain accurate geometries and energies for Mg^{2+} ligand binding and the formation of di-metal Mg^{2+} complexes containing a bridging hydroxide. Current non-bonded parameters exist for Mg^{2+} in the CHARMM27 force field[301] and OH^- parameters are available from Lee *et al.*[317] However, these parameters are not optimal for the specific problems that involve di-metal bridges or RNA phosphate binding. Hence, the parameters for OH^- and Mg^{2+} have been re-optimized specifically for these types of systems.

DFT calculations have been run on several model systems[129] which formed the training set for the Mg^{2+} and OH^- parameterization. Each structure of the training set (Figure 5.5) contained one or two Mg^{2+} ions with a first coordination sphere combination of H_2O, OH^- and/or DMP^- ligands. In the structures with two Mg^{2+} ions present, a OH^- ion was placed in a μ-bridging position between the two metal ions. CHelpG charges were calculated for OH^- and the L-J parameters for Mg^{2+} and OH^- were iteratively adjusted until the best fit with geometry and water binding energy was obtained (Table 5.1). Mg^{2+} coordination distances were significantly improved with the new force field parameters for ligand binding with H_2O, OH^- and DMP^-. This improvement was most pronounced for the di-metal bridging hydroxide structures which lengthen the $Mg^{2+}\cdots OH^-$ interaction by almost 0.30 Å and contract the $Mg^{2+}\cdots OH\cdots Mg^{2+}$ angle by almost 10 degrees. The new parameters also improved the Mg^{2+} water

binding energy by approximately 1.0 kcal/mol. In the current study, only complexes that involve inner-sphere coordination of Mg^{2+} were used in the parameterization, and transferability to the important case of outer-sphere coordination remains to be tested with simulation.

2'O Deprotonated Sugar Ring

The first step of transesterification for a RNA phosphate is deprotonation of the sugar 2'OH. This creates a nucleophile of sufficient strength to attack the phosphorous of an adjacent or nearby phosphate group.[273] A "patch"[262] has been prepared for the standard CHARMM27 RNA nucleotide which removes the 2'OH hydrogen and modifies the $2'O^-$ charge, L-J interactions and geometry of nearby atoms. Parameterization of this patch is discussed below.

Table 5.1: Mg^{2+} and OH^- parameter fitting results

Bond (Å)	**CHARMM**[1]	**DFT**	**CHARMM***
HT–OW	0.97 (0.00)	0.97	0.97 (0.00)
MG· · ·OH2	1.99 (-0.13)	2.12	2.13 (0.01)
MG· · ·OW	1.81 (-0.19)	2.00	2.00 (0.00)
MG· · ·ON3	1.86 (-0.17)	2.03	2.01 (-0.02)
MG· · ·MG	3.56 (-0.33)	3.89	3.91 (0.02)
Angle (°)	**CHARMM**[1]	**DFT**	**CHARMM***
MG· · ·OW· · ·MG	155.8 (7.6)	148.2	152.7 (4.5)
Energy (kcal/mol)	**CHARMM**[1]	**DFT**	**CHARMM***
$(Mg(H_20)_5 + H_2O) \rightarrow Mg(H_20)_6$	-27.4 (0.0)	-27.4	-28.3 (-0.9)

[1]Training set (see Figure 5.5) average errors for geometry and reaction energy fitting. With the exception of the hydroxide OW and HT atom types from Lee *et. al.*,[317] the remaining parameters were originally from the standard CHARMM27 force field Foloppe *et. al.*[262] OH2 is the CHARMM27 water oxygen and ON3 is the non-bridging oxygen of DMP^-. See the Supporting Information for more data on the fitting function weight values used.

Table 5.2: CHelpG charge fitting for deprotonated ribose phosphate

Atom	$\mathbf{q_o}$	$\mathbf{q_P}$	$\mathbf{q'_P}$	$\mathbf{q_D}$	$\mathbf{q'_D}$	$\delta\mathbf{q}$	$\mathbf{q^*_D}$
C1'	0.07	0.19	0.00	0.31	-0.19	**-0.19**	-0.12
H1'	0.09	0.03	0.09	-0.12	0.09	0.00	0.09
H1"	0.09	-0.04	0.09	-0.20	0.09	0.00	0.09
C2'	0.14	0.35	0.21	0.70	0.30	**0.09**	0.23
H2"	0.09	-0.05	0.09	-0.31	0.09	0.00	0.09
O2'	-0.66	-0.71	-0.72	-0.97	-0.97	**-0.25**	-0.91
H2 '	0.43	0.42	0.43	-	-	**-0.43**	-
C3'	0.01	0.30	0.20	0.65	0.37	**0.17**	0.18
H3'	0.09	-0.01	0.09	-0.19	0.09	0.00	0.09
O3'	-0.57	-0.57	-0.57	-0.63	-0.63	**-0.06**	-0.63
P	1.50	1.23	1.23	1.37	1.37	**0.14**	1.64
O1P	-0.78	-0.77	-0.77	-0.84	-0.83	**-0.06**	-0.84
O2P	-0.78	-0.76	-0.76	-0.81	-0.83	**-0.06**	-0.84
O3P	-0.57	-0.50	-0.50	-0.58	-0.58	**-0.08**	-0.65
C3T	-0.17	0.31	-0.10	0.36	-0.16	**-0.06**	-0.23
H3T1	0.09	-0.06	0.09	-0.08	0.09	0.00	0.09
H3T2	0.09	-0.05	0.09	-0.08	0.09	0.00	0.09
H3T3	0.09	-0.03	0.09	-0.09	0.09	0.00	0.09
C4'	0.07	0.27	0.06	0.14	-0.12	**-0.18**	-0.11
H4'	0.09	-0.03	0.09	-0.07	0.09	0.00	0.09
H4"	0.09	0.00	0.09	-0.01	0.09	0.00	0.09
O4'	-0.50	-0.52	-0.52	-0.55	-0.55	**-0.03**	-0.53

Adjusted charges ($\mathbf{q'_P/q'_D}$) were determined by changing all CHelpG hydrogen charges to be consistent with CHARMM27 by the procedure outlined in Equations 5.3 and 5.4 in the text. This difference between the CHARMM27 ($\mathbf{q_o}$) and CHelpG charges ($\mathbf{q_P/q_D}$) was then added into the nearest heavy atom charge. Once the adjusted atomic charge differences between the protonated and deprotonated structures were determined ($\delta\mathbf{q}$), these differences were then added to the standard CHARMM27 charges to correct for deprotonation ($\mathbf{q^*_D}$).

To derive new charges for a deprotonated nucleotide residue, a ribose methyl phosphate in its protonated and deprotonated states was optimized using DFT. It was found that an "in-line" geometry minimum did not exist in the gas phase without the solvated ribozyme environment. In order to parameterize a deprotonated ribose methyl phosphate (ribose MeP^-) residue to a structure that models the in-line phosphate attack conformation, a minimization was performed with the corresponding α, ϵ and ζ dihedrals of the phosphate-sugar backbone fixed to the minimum energy geometry found for

the phosphorane structure (Figure 5.4). In this way, the backbone of the ribose MeP^- was constrained to be "in-line". CHelpG charges, constrained to reproduce molecular dipole moment for neutral molecules, were then calculated for the atoms of the optimized structures (Table 5.2). Since the original CHARMM27 charges (q_o) were obtained via a slightly different procedure than those of the present DFT/CHelpG method (q), adjustments to the CHelpG charges (q') were made to create charges more consistent with the constrained CHARMM27 force field parameterization. For instance, in the CHARMM27 force field, all non-polar hydrogens are constrained to have a charge of 0.09 e while polar hydroxyl hydrogens are assigned a charge of 0.43 e. The calculated CHelpG charges were adjusted for this by changing the value of the hydrogens to be consistent with the CHARMM27 force field while adding the difference created from this adjustment to the adjoining heavy atoms (Table 5.2). The differences between the adjusted CHelpG charges for the protonated and deprotonated ribose methyl phosphate were then calculated and used to modify the original CHARMM27 charges of a neutral nucleotide as follows:

$$\delta q = q'_P - q'_D \tag{5.3}$$

$$q^*_D = q_o + \delta q \tag{5.4}$$

where q'_P and q'_D are the adjusted DFT/CHelpG charges for the protonated and deprotonated structures, respectively, q_o are the CHARMM27 charges, and q^*_D are the new charges for the deprotonated structure. Final partial atomic charge parameters are shown in Table 5.2. It should be noted that constraints on the charge fitting to create unit charge groups within residues was not used in this method.

Once charges were calculated for the deprotonated ribose 2'O^- oxygen, L-J parameters were determined by fitting to the 2'O^- water coordination distance (1.713 Å) and interaction energy (-18.4 kcal/mol) obtained from DFT optimization of a ribose MeP^- (Table 5.3). Bond and angle parameters were fit to the differences between the gas-phase DFT geometry of protonated and ribose MeP^- (Table 5.4). Force constants were predicted by fitting to relaxed potential energy surface scans. All dihedrals except for H3'-C3'-C2'-O2' were set to zero in the original protonated CHARMM27 structure and remained so in the new parameterization. The H3'-C3'-C2'-O2', which has a K_χ of 0.195 kcal/mol/radian^2 and 0.0° phase for the 3-fold term, was retained. Final fitting was performed for the residue patch within the RNA sequence UCA taken from the

Table 5.3: RNA Lennard-Jones parameters

Parameter	Protonated Methyl Ribose	Deprotonated Methyl Ribose	Phosphorane Methyl Ribose	2',3'-cyclic Methyl Ribose
$R_{min}/2_{O2'}$	1.77	1.75	1.76	1.77
$\epsilon_{O2'}$	-0.1521	-0.3236	-0.2378	-0.1521

L-J parameters for the deprotonated ribose phosphate were fit to the water binding coordination distance (1.713 Å) and the water binding coordination energy (18.4 kcal/mol) with TIP3P water (See Table 5.4 as well) . L-J parameters for axial phosphorane oxygens (bond order ≈ 0.5) were determined by averaging between the protonated (bond order ≈ 0.0) and 2',3'-cyclic (bond order ≈ 1.0) structures.

"early intermediate" active site geometry in the x-ray crystal structure of a hammerhead ribozyme.[268] ABNR minimization was performed for this sequence with the U and A residues fixed and, where necessary to improve fitting, small adjustments to the internal parameters were made manually. The final results (Table 5.4) were able to reproduce the DFT structure and water binding energy very closely.

Table 5.4: Geometry fitting results of deprotonated ribose phosphate

Bond (Å)	Prot[a] DFT	Deprot[b] DFT	Relative Difference	Prot[a] CHARMM	Adjusted CHARMM*
O_2'-C_2'	1.43	1.33	-0.10	1.43	1.33 (0.00)
O_2'-C_2'-H_2'	109.6	115.8	6.2	109.9	116.1 (0.0)
O_2'-C_2'-C_1'	110.2	112.7	2.5	114.0	116.5 (0.0)
O_2'-C_2'-C_3'	109.8	116.4	6.6	112.2	118.8 (0.0)

Reaction	Coordination Distance DFT	Coordination Distance CHARMM*	Coordination Energy DFT	Coordination Energy CHARMM*
ribose MeP^- + H_2O → ribose MeP^-:H_2O (kcal/mol)	1.71	1.70 (0.00)	-18.4	-18.4 (0.0)

Where ribose MeP^- stands for the deprotonated ribose methyl phosphate. Final fitting for the Adjusted CHARMM* residue was performed within the RNA sequence active site taken from the "early intermediate" hammerhead ribozyme x-ray crystal structure.[268] [a]Protonated ribose. [b]Deprotonated ribose.

Phosphate Transition State Analog

The deprotonated ribose 2'O^- attack at the phosphate produces a phosphorane that is a transition state in the gas phase (Figure 5.4). Partial charges for the phosphorane atoms were determined similarly to those for the 2'O^- ribose MeP^- described above. However, due to the instability of dianionic phosphorane in the gas phase, DFT optimization was carried out with the axial P-O2' bond length fixed to the P-O2' bond length of a 2'O^- ribose methyl phosphorane (ribose $MePA^{2-}$) optimized in the aqueous phase (1.986 Å) using the Polarizable Continuum Model (PCM) solvation model.[224,318] Comparison of similar neutral and anionic phosphorane structures optimized in both the gas and aqueous phases, suggests that the error for this constraint should be less than 0.02 Å for all but the axial P-O bonds which may have errors as high as 0.07 Å. Calculated CHelpG charges for hydrogen atoms were adjusted to 0.09 e with heavy atoms taking up the difference as before (Table 5.5). Final charges (q^*_{TS}) for the phosphorane structure were determined by adding the adjusted CHelpG charge difference (δq) between the ribose methyl phosphorane and the 2'O^- ribose MeP^- to the previously determined CHARMM* charges (q^*_D) for the the 2'O^- ribose MeP^- (see Table 5.5).

Bond length and angle equilibrium parameters for the ribose $MePA^{2-}$ were determined from fitting to the gas phase DFT optimization of the partially frozen dianionic ribose $MePA^{2-}$ structure described previously. Force constants for bonds and angles were calculated by fitting to relaxed potential energy surface (PES) scans for each P-O bond and O-P-O angle of a monoanionic phosphorane (which is stable in the gas phase). The final ribose $MePA^{2-}$ residue was fit within the RNA sequence UCA taken from the near in-line geometry in the x-ray crystal structure of a hammerhead ribozyme "late-intermediate".[319] ABNR minimization was performed for this sequence with the U and A residues fixed and where necessary, small adjustments to the internal parameters were made to improve fitting. Because the penta-coordinated geometry of the phosphorane is quite rigid, the dihedrals had negligible effect, and hence were set to zero. The final fitting results matched with the DFT results almost exactly (Tables 5.6).

Table 5.5: CHelpG charge fitting for ribose phosphorane

Atom	$\mathbf{q^*_D}$	$\mathbf{q_D}$	$\mathbf{q'_D}$	$\mathbf{q_{TS}}$	$\mathbf{q'_{TS}}$	$\delta\mathbf{q}$	$\mathbf{q^*_{TS}}$
C1'	-0.12	0.31	-0.19	0.21	-0.09	**0.10**	-0.02
H1'	0.09	-0.12	0.09	-0.10	0.09	0.00	0.09
H1"	0.09	-0.20	0.09	-0.02	0.09	0.00	0.09
C2'	0.23	0.70	0.30	0.42	0.17	**-0.13**	0.10
H2"	0.09	-0.31	0.09	-0.16	0.09	0.00	0.09
O2'	-0.91	-0.97	-0.97	-0.74	-0.74	**0.23**	-0.68
C3'	0.18	0.65	0.37	0.54	0.29	**-0.08**	0.10
H3'	0.09	-0.19	0.09	-0.16	0.09	0.00	0.09
O3'	-0.63	-0.63	-0.63	-0.62	-0.62	**0.01**	-0.62
P	1.64	1.37	1.37	1.54	1.54	**0.17**	1.81
O1P	-0.84	-0.84	-0.83	-0.88	-0.90	**-0.08**	-0.92
O2P	-0.84	-0.81	-0.83	-0.92	-0.90	**-0.08**	-0.92
O3P	-0.65	-0.58	-0.58	-0.68	-0.68	**-0.10**	-0.75
C3T	-0.23	0.36	-0.16	0.54	-0.21	**-0.05**	-0.28
H3T1	0.09	-0.08	0.09	-0.14	0.09	0.00	0.09
H3T2	0.09	-0.08	0.09	-0.16	0.09	0.00	0.09
H3T3	0.09	-0.09	0.09	-0.18	0.09	0.00	0.09
C4'	-0.11	0.14	-0.12	0.02	-0.14	**-0.02**	-0.13
H4'	0.09	-0.07	0.09	-0.04	0.09	0.00	0.09
H4"	0.09	-0.01	0.09	0.02	0.09	0.00	0.09
O4'	-0.53	-0.55	-0.55	-0.53	-0.53	**0.02**	-0.51

CHelpG adjusted charges for determining ribose phosphorane partial atomic charges. New charges were created by taking the adjusted CHelpG charge difference $\delta\mathbf{q}$ between the ribose MeP^- charges calculated in Table 2 and the ribose phosphorane. $\mathbf{q_{TS}}$ are the ribose phosphorane CHelpG charges, $\mathbf{q'_{TS}}$ are the adjusted CHelpG charges and $\mathbf{q^*_{TS}}$ are the final CHARMM* charges. See Equations 5.3 and 5.4 of text for further details.

2',3'-Cyclic Phosphate

Transesterification terminates in the creation of a 2',3'-cyclic phosphate after exocyclic P-O5' bond cleavage. The 2',3'-cyclic phosphate is similar to a standard RNA ribose methyl phosphate residue except for the cyclization of the O2'-P-O3'-C3'-C2' atoms (see Figure 5.4 and the parameter file in the Supporting Information). In the cyclic residue, the O2' is bonded to the phosphorus in the same manner as the O3'. Due

Table 5.6: Geometry fitting results for ribose phosphorane parameterization

Bond (Å)	**DFT**	**CHARMM***
C_2-O_2	1.361	1.361 (0.00)
P-O_2	1.99	1.99 (0.00)
P-O_3	1.79	1.79 (0.00)
P-O_5	1.84	1.84 (0.00)
P-$O_{R/S}$	1.53	1.53 (0.00)
O_2-O_5	3.78	3.79 (0.01)
Angle (°)	**DFT**	**CHARMM***
C_1-C_2-O_2	113.5	113.7 (0.2)
C_3-C_2-O_2	105.9	105.9 (0.0)
C_2-O_2-P	109.3	109.3 (0.0)
C_3-O_3-P	117.5	117.5 (0.0)
O_2-P-O_5	163.5	164.0 (0.5)
O_R-P-O_S	129.1	129.1 (0.0)
O_3-P-$O_{R/S}$	115.3	115.3 (0.0)
O_2-P-$O_{R/S}$	91.4	91.4 (0.0)
O_5-P-$O_{R/S}$	95.4	95.3 (0.1)
O_2-P-O_3	81.0	81.0 (0.0)
O_3-P-O_5	83.3	83.3 (0.0)

Final fitting for the Adjusted CHARMM* residue was performed within the RNA sequence active site taken from the "late intermediate" hammerhead ribozyme x-ray crystal structure.[319]

to the near symmetric equivalence between the O2' and O3' cyclic ribose phosphate sequences, the bond length, angle, and dihedral parameters are assumed to be identical. Therefore, O2' was assigned the same atom type as O3' (type ON2). Adjusted CHelpG charges for the 2',3'-cyclic phosphate ring were calculated and force constants were determined through fitting to relaxed PES scans. Parameters were fit to the DFT data using ABNR minimization of the CHARMM* residue. New dihedrals were set to zero analogous those of the CHARMM27 ribose ring. Final parameters fit with very small errors with respect to the DFT results (Table 5.7).

Table 5.7: Geometry fitting results for 2',3'-cyclic phosphate parameterization

Bond (Å)	**DFT**	**CHARMM***
O2'-P	1.71	1.71 (0.00)
O3'-P	1.71	1.71 (0.00)
P-O1P	1.50	1.50 (0.00)
P-O2P	1.50	1.50 (0.00)
Angle (°)	**DFT**	**CHARMM***
C2'-O2'-PC	109.4	109.4 (0.0)
C3'-O3'-PC	109.4	109.4 (0.0)
O2'-PC-O3'	91.0	91.0 (0.0)
O2'-PC-O1/2C	109.0	109.0 (0.0)
O3'-PC-O1/2C	109.0	109.0 (0.0)
O1C-PC-O2C	124.5	124.5 (0.0)

Training set average geometries for geometry fitting of 2',3'-cyclic phosphate.

Thio-Substitutions

To determine the role of divalent metal ions in a reaction, it is often useful to perform thio effect experiments.[175,320] Changing the oxygen atom to a sulfur effects how strongly a divalent metal can bind to this position due to the differences between their sizes, polarizabilities, and bond lengths with phosphorous. While a "hard" ion like Mg^{2+} will bind very tightly to a correspondingly "hard" oxygen, it will bind much more weakly to a "softer", more diffuse sulfur.[321] If the divalent metal ion binding at a certain position is required for catalysis, a large decrease in the reaction rate should be seen upon thio-substitution at this position. However, the interpretation of thio affect results are sometimes inconclusive[278,280,322,323] especially when possible conformational changes induced by the sulfur are considered. Molecular simulation of thio-substituted phosphates and phosphoranes and their role on structure would aid in the interpretation of experimental thio effects.

A training set of phosphate compounds, bound and unbound to hexa-coordinated Mg^{2+}, with both single and double sulfur substitutions was constructed (Figure 5.6). Adjusted CHelpG partial atomic charge calculations were calculated for both mono- and di-thio DMP^{-} substitutions. Iterative optimization of the L-J parameters for sulfur

was made until the best possible fit to the DFT calculated sulfur water-binding distance and relative water-binding energy between DMP^- and di-thio substituted DMP^- was obtained. Force constants and equilibrium geometry parameters were then optimized iteratively along with the L-J parameters until minimization of the χ^2 function was achieved.

In order to improve the differential binding of Mg^{2+} to the non-bridging phosphoryl O and S atoms, a NBFIX term[262] was used to parameterize specifically for Mg^{2+}-S interactions; i.e., a specific value for the 2-body non-bonded van der Waals parameters was included explicitly for Mg^{2+}-S rather than using the Lorentz-Berthelot combining rules to derive the 2-body parameters from 1-body Mg^{2+} and S parameters. The resulting fit was further improved by the introduction of three new atom types. Two for the non-bridging oxygen (ONS) and sulfur (SO) of a mono-substituted phosphate and one for the non-bridging sulfur (SS) of a di-thiosubstituted phosphate. The DFT training set average binding distance for a Mg^{2+}-S was 2.55 Å as compared to the CHARMM* average value of 2.06 Å (Table 5.8). The shortened distances arose out of the need to strengthen the binding energy results that, in the absence of explicit polarization on the soft sulfur atoms, is considerably underestimated with non-polarizable force fields. In the case of water binding to the substituted sulfurs, the distances are in better agreement and the relative binding energies are all within 1.0 kcal/mol of the DFT values.

The ζ *ab initio* torsional potential energy surfaces for the native and non-bridging thio-substituted DMP^- are shown in Figure 5.3 (top frame). The minima identified agree well with fully relaxed structures in previous work by Florian *et al.*[324] A significant difference between the thio-substituted and unsubstituted surfaces is the minima at the *staggered* position, 180°, which is negligible for the native DMP^-, but of almost equal energy to the *gauche* minima in the case of dithio-substituted DMP^-. Also noteworthy is the shift of the *gauche* minima (70°) and *gauche-staggered* barrier with thio substitution. New torsional parameters, summarized in Table 5.9, are defined along the S-P-O-C and O-P-O-C dihedrals to qualitatively reproduce the changes indicated by the *ab initio* calculations, while maintaining consistency with the CHARMM27 force field.

A scan of the DMP^- and thio-substituted torsions with the same constraints as the *ab initio* calculations using the CHARMM27 parameters and those developed in this

Table 5.8: Geometry and binding energy results for phosphate thio-substitution parameterization

Bond (Å)	**Relative DFT**	**CHARMM***
SO$\cdots$OT	2.65	2.41 (-0.24)
SS$\cdots$OT	2.65	2.34 (-0.31)
SO$\cdots$MG	2.56	2.06 (-0.50)
SS$\cdots$MG	2.54	2.05 (-0.49)
SO-P	2.01	2.00 (-0.01)
SS-P	2.00	2.00 (0.00)
Angle (°)	**DFT**	**CHARMM***
ON2-P-SO	107.9	107.9 (0.0)
ON2-P-SS	108.4	108.0 (-0.4)
ONS-P-SO	121.0	120.6 (-0.4)
SS-P-SS	121.4	121.1 (-0.3)
Binding Energy (kcal/mol)	**Relative DFT**	**CHARMM***
DMP(oo)/(so)$\cdots H_2O$	-2.7	-2.8 (-0.1)
DMP(oo)/(ss)$\cdots H_2O$	-3.4	-3.8 (-0.4)
DMP(oo)/(ss)-(g-t)/(g-g)	-1.8	-1.5 (0.2)
Mg(HOH)5(DMP)-(oo)/(so)$\cdots H_2O$	-4.5	-5.2 (-0.7)
Mg(HOH)5(DMP)-(oo)/(ss)$\cdots H_2O$	-6.8	-7.5 (-0.7)

Training set average geometries for geometry and reaction energy for L-J fitting. DMP-oo/ss$\cdots H_2O$ above is the relative energy difference between water binding to DMP^- with and without thio-substitution at the non-bridging oxygens. The binding distance for SO$\cdots$MG and SS$\cdots$MG were fit using a nonbond fix (NBFIX). Geometries were fit to the shift predicted by DFT relative to the unsubstituted DMP^-. (g-t) vs. (g-g) indicates the *gauche-trans* vs. *gauche-gauche* conformations.

work are shown in Figure 5.3 (bottom frame). As indicated in Foloppe *et al.*,[262] the barrier between *gauche-gauche* states in the unsubstituted DMP^- has been significantly lowered to reproduce results from experiment, while the *gauche-staggered* barrier is left relatively unchanged. The thio-substituted structures retain the lowered *gauche-gauche* barrier and the *gauche* minima is shifted as indicated by the *ab initio* calculations. The *staggered* conformation has been lowered in energy for both thio-substituted molecules. For the singly substituted case, the minima at 300° was also lowered in addition to the minima at 60°, but the relative energy of the two gauche and eclipsed conformations were maintained. The lowering *staggered* conformation may be of particular importance to catalytic differences in native and thio-substituted structures since conformational

Table 5.9: Dihedral parameters of non-bridging oxygen and sulfur for thio-substituted phosphate

	Atom Types				Force Constant	Fold	Phase
Thio-substitution	ONS	P	ON2	CN9	0.20	2	0.00
	ONS	P	ON2	CN9	0.20	3	0.00
	SO	P	ON2	CN9	0.40	1	180.00
	SO	P	ON2	CN9	0.10	3	0.00
Dithio-substitution	SS	P	ON2	CN9	0.45	1	180.00
	SS	P	ON2	CN9	0.20	2	0.00
	SS	P	ON2	CN9	0.10	3	0.00

deformation is a prerequisite to reaction.[289,310,325,326]

Table 5.10: Geometry fitting results for thio-substituted ribose phosphorane parameterization

Bond (Å)	**DFT**	**CHARMM***
P-$S_{R/S}$	2.10	2.10 (0.00)
Angle (°)	**DFT**	**CHARMM***
O_R-P-S_S	125.4	125.3 (0.1)
S_R-P-S_S	125.3	125.3 (0.0)
O_3-P-$S_{R/S}$	113.4	113.5 (0.1)
O_2-P-$S_{R/S}$	91.4	91.4 (0.0)
O_5-P-$S_{R/S}$	97.3	97.3 (0.0)
ribose MePA-(oo/os)$^{2-}$ + H_2O → ribose MePA-(oo/os)$^{2-}$:H_2O (kcal/mol)	6.6	7.9 (1.3)
ribose MePA-(oo/ss)$^{2-}$ + H_2O → ribose MePA-(oo/ss)$^{2-}$:H_2O (kcal/mol)	6.4	7.7 (1.3)

Training set average geometries and energies for geometry fitting and L-J fitting of dianionic ribose methyl phosphorane.

L-J parameters for the non-bridging atoms of the penta-coordinated di-anionic phosphorane structures were parameterized using a training set which contained methyl ribose phosphorane (ribose MePA^{2+}), mono-thio-substituted ribose MePA-so^{2+} and di-thio-substituted ribose MePA-ss^{2+}, with and without water bound (see Figure 5.7). In the case of the unsubstituted phosphorane, it was required to freeze the P-O2' bond

as discussed above. Stable gas-phase sulfur substituted phosphoranes were used for the other structures. Both the non-bridging S and O atoms were reparameterized in this case due to the large differences in geometry and charge distribution of a phosphate and phosphorane. As before, equilibrium geometry values were fit to DFT structures with force constants for all modified bonds and angles determined through fitting to relaxed PES scans. Geometry fitting results are shown in Table 5.10.

5.1.4 Conclusion

Molecular simulation force field parameters are presented for a series of non-standard residues important in RNA catalysis. Parameters are based on density-functional calculations and developed to be consistent with the CHARMM27 all-atom empirical force field for nucleic acids. Parameters have been developed for an activated (2'O deprotonated) ribose phosphate representing an early reactive state, a penta-coordinate phosphorane intermediate/transition state model, and a 2',3'-cyclic phosphate transesterification product. In addition, parameters for thio-substituted analogs important in the study of experimental thio effects, and specific single and di-metal Mg^{2+} complexes and μ-bridging OH^- ion parameters have been developed. These parameters, which are optimized specifically for the present systems, will allow the simulation of reactive intermediates and experimentally modified residues important in the study of ribozyme mechanisms. Further validation and testing of these parameters in simulations is important, although this will likely be a lengthy and tedious endeavor due to the lack of available experimental structural data for the reactive intermediates and thio-substituted nucleic acids. Ultimately, it is the hope that simulations of these systems, together with experiment, will help paint a more detailed picture of the catalytic mechanisms of RNA catalysis.

5.1.5 Additional Figures

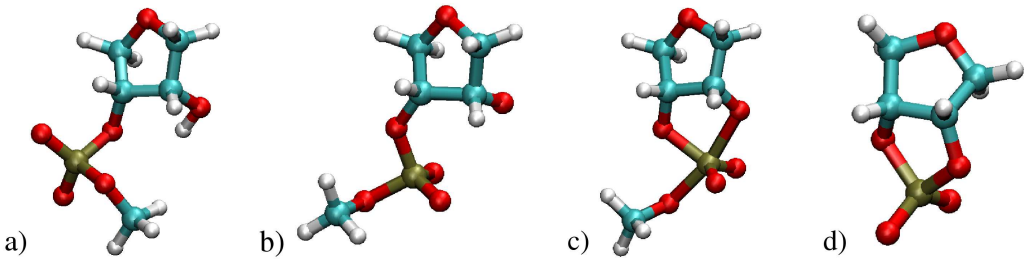

Figure 5.4: CHARMM27 standard and modified nucleotide residues: a) A standard CHARMM27 ribonucleotide analog (no base) with the 2' position protonated; b) 2' deprotonated ribose phosphate (i.e., activated at the 2' position); c) ribose phosphorane (i.e., intermediate/transition state model); d) 2',3'-cyclic ribose phosphate (i.e., transesterification product). Color coding of atoms is as follows: carbon=turquoise, oxygen=red, phosphorous=brown, hydrogen=white.

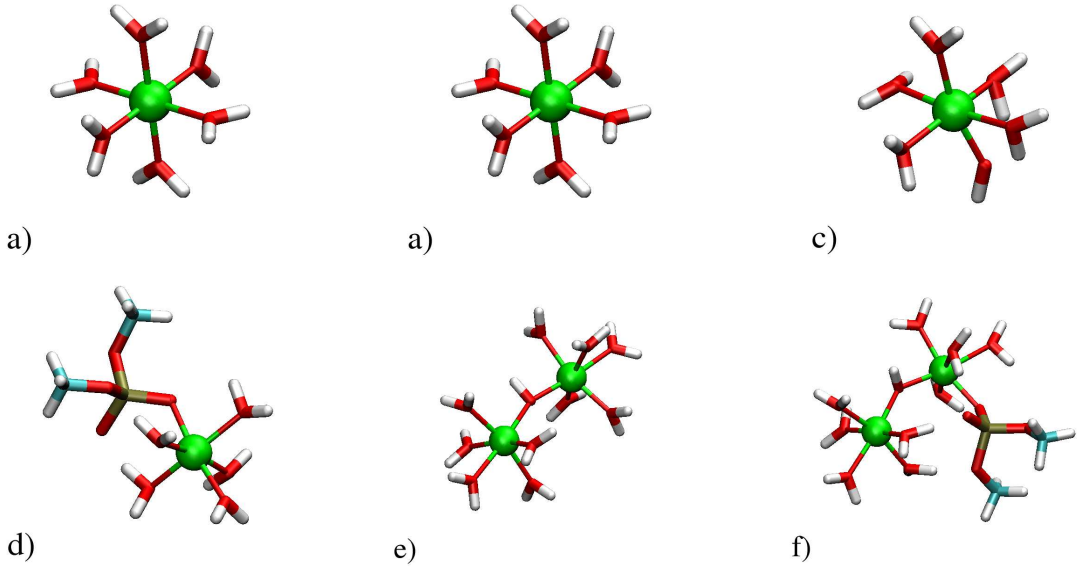

Figure 5.5: Complexes used for Mg^{2+} and OH^{-} Parameterization a) $Mg(H_2O)_6^{2+}$, b) $Mg(H_2O)_5^{2+}$, c) $Mg(H_2O)_5(OH^{-})^{1+}$, d) $Mg(H_2O)_5$:DMP^{1+}, e) $Mg(H_2O)_5{\cdot}(OH^{-}){\cdot}Mg(H_2O)_5^{3+}$, f) $Mg(H_2O)_5{\cdot}$ $(OH^{-}){\cdot}Mg(H_2O)_4$:$DMP^{2+}$. Color coding of atoms is as follows: magnesium=green, oxygen=red, carbon=turquoise, hydrogen=white.

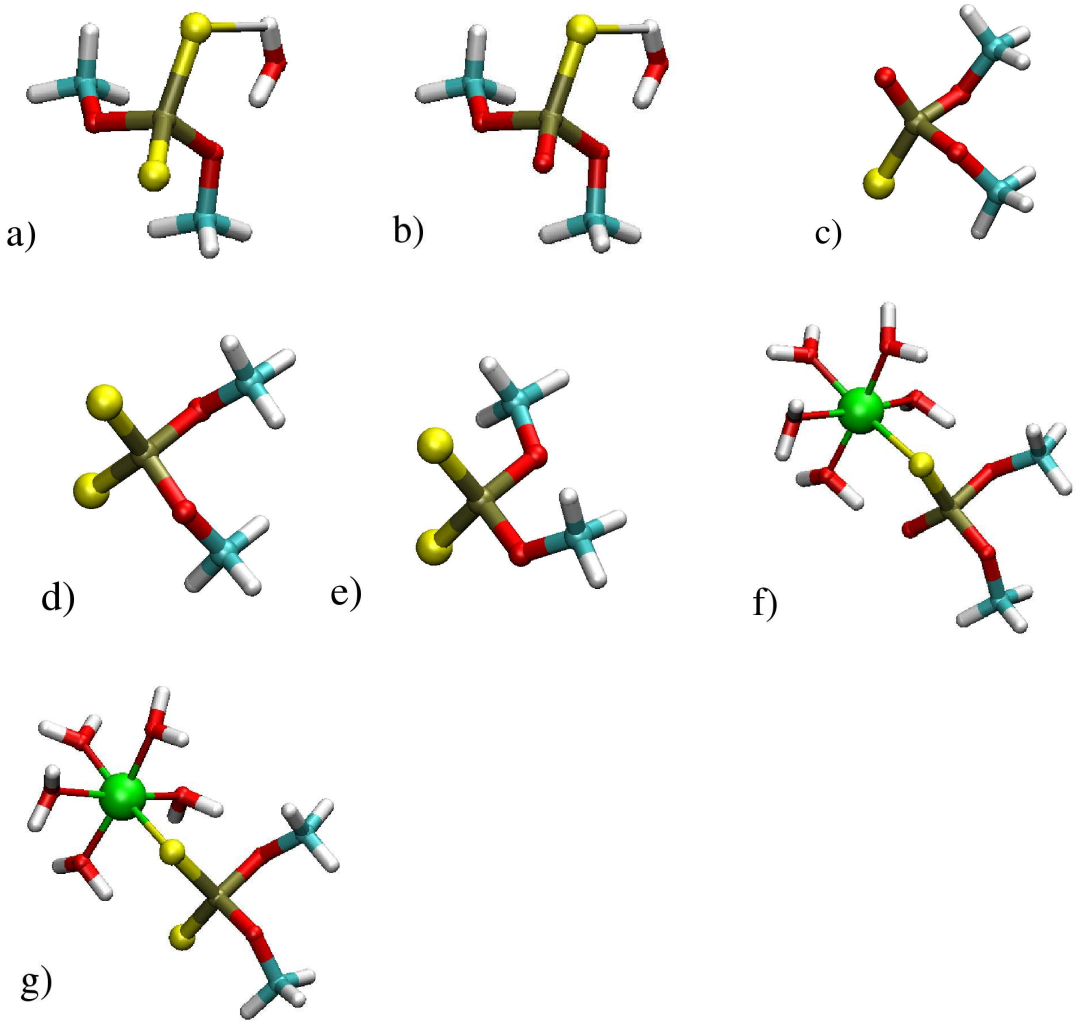

Figure 5.6: Complexes used for S and Mg^{2+} Parameterization a)DMP-ss^{-}:HOH, b)DMP-so^{-}:HOH, c)DMP-so^{-}, d)DMP-$ss^{-}{}_{g-g}$, e)DMP-$ss^{-}{}_{g-t}$, f)DMP-so^{-}:$[Mg(HOH)_5]^{2+}$, g)DMP-ss^{-}:$[Mg(HOH)_5]^{2+}$. Color coding of atoms is as follows: magnesium=green, oxygen=red, carbon=turquoise, hydrogen=white, phosphorous=brown, sulfur=yellow.

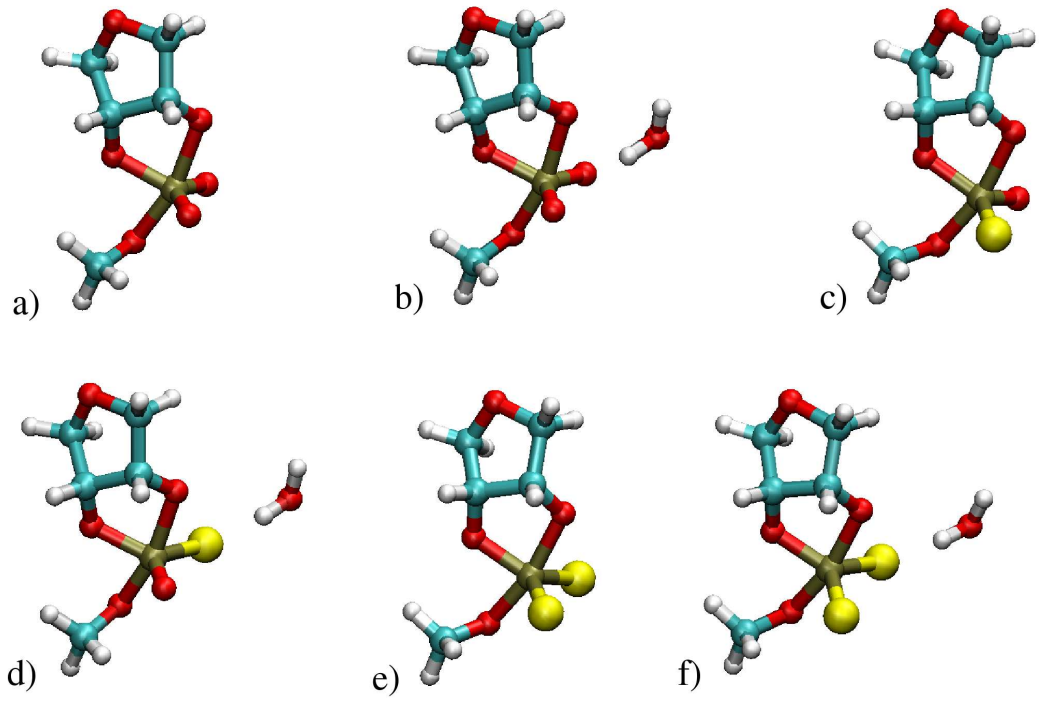

Figure 5.7: Complexes used for Thio-substituted Phosphorane Parameterization a)ribose MePA^{2-}, b)ribose MePA^{2-}:HOH, c)ribose MePA-so^{2-}, d)ribose MePA-so^{2-}:HOH, e)ribose MePA-ss^{2-}, f)ribose MePA-ss^{-}:HOH. Color coding of atoms is as follows: oxygen=red, carbon=turquoise, hydrogen=white, phosphorous=brown, sulfur=yellow.

5.2 Solvation of phosphoric acid

The following section is not itself parameterization, but rather an example of the importance of having accurate data to parameterize against. This work is currently unpublished.

5.2.1 Introduction

An accurate description of solvation is critical for theoretical calculations of biological molecules and their reactions. Solvation has a considerable effect on the electronic structure, and thereby affects the equilibrium geometry and dynamic properties of solutes.[75–77] In Chapter 2, the importance of having accurate gas-phase basicities (GPB) was discussed with one application being the prediction of pK_{a} values using a thermodynamic cycle (as in Figure 2.1 in Section 2.2.2). The other key component for pK_{a} prediction is the solvation free energy ($\Delta \mathrm{G}_{solv}$) for the acid, conjugate base, and proton. Implicit solvation models afford a computationally tractable means to estimate these values and have been under development for almost a century.[327]

Obtaining multiple pK_{a} values[103] requires the accurate prediction of the relative solvation free energy between the acid and its conjugate base, which often involves highly charged species making it particularly challenging for computational models. Among the many possible explanations why current models cannot attain the accuracy needed for absolute pK_{a} prediction may be the invariance of the solute cavity, which differentiates the system into a solute and solvent, with respect to solute charge state. Qualitatively speaking, the solvent should pack more closely to the solute as it becomes more ionic because of the favorable Coulombic interaction between the permanent dipole of the solvent, in the case of water, and the charge distribution of the solute. Models that separate the solute and solvent by a static boundary are not directly able to address this effect.

Phosphoric acid and its derivatives (e.g. DNA backbone, ATP, etc) are ubiquitous in biochemistry and a challenging benchmark compound for pK_{a} prediction and experiment.[328] Due to having three titratable protons on four possible sites, phosphoric acid solvation is very complex with intramolecular hydrogen bonding, symmetry, multiple species present at a given pH, and highly charged anions all are necessary considerations.

For these reasons, while the pK_{a} values are known for all three deprotonations and the GPB can be reasonably estimated, the solvation free energies of phosphoric acid and its anions are more ambiguous. The first step toward creating models capable of dealing with such a complicated systems is having reliable reference data to test against. This section presents both an investigation into the currently used solvation free energies for phosphoric acid and its anions as well as a benchmark of common implicit solvation models.

5.2.2 Methods

Electronic structure and solvation calculations

All electronic structure, thermochemical analysis, and solvation calculations were performed using the Gaussian03 suite of programs,[122] except for the coupled cluster calculations for W1 which were run with MOLPRO[329] and the SM5.43 and SM6 solvation calculations run with Gaussian98[330] and the MNGSM module.[331]

Solvation calculations were performed using the CPCM,[225] IEFPCM,[332] SCIPCM,[333] COSMO,[82] SM5.43,[334] and SM6[87] implicit solvation models using the default solvation radii or isodensity value. Single point solvation calculations were performed on the B3LYP/cc-pVTZ+d geometries from the W1 calculations. Self-consistent implicit solvation optimizations were also calculated at the same level of theory and basis set. SM5.43 does not support the cc-pVTZ+d basis set so the 6-31G(d) basis set, the largest supported basis set, was employed.

Calculation gas-phase basicities

The gas-phase basicity (GPB) of a species A^- are related to the gas-phase reaction:

$$A^-_{(g)} + H^+_{(g)} \rightarrow AH_{(g)} \tag{5.5}$$

The GPB of A^- is defined as the negative of the Gibbs free energy change (ΔG) of the process in Eq. 5.5.[116] Here, the A^- is the conjugate base associated with the neutral acid AH.

These quantities have been determined for a variety of molecules,[121] but when experiment is not available accurate quantum methods can be used. Many *ab initio* multilevel and density functional methods have been benchmarked on important phosphate species[139,140] (as in Chapter 2).

Experimental solvation free energies

Solvation free energies for a range of molecules, both neutral and ionic, have been organized by many researchers.[102,334–338] Values for neutral species can be obtained experimentally from Henry's law constants or the vapor pressure of the pure substance and the activity coefficient.[339] For ionic species a thermodynamic cycle is often used;[104,105,112,114,336,337,340] Figure 2.1 is a commonly used cycle.

To determine the solvation energy for anions, this cycle requires the GPB, aqueous reaction free energy, solvation free energy of the neutral species, and the solvation free energy of the proton, and is related as

$$\Delta G^{\circ}_{solv}\left(A^{-}\right) = \triangle G^{\circ}_{aq} - \Delta G^{\circ}_{gas} + \Delta G^{\circ}_{solv}\left(HA\right) - \Delta G^{\circ}_{solv}\left(H^{+}\right) \tag{5.6}$$

The aqueous reaction free energies can be obtained from pK_{a} values through the relation

$$\Delta G^{\circ}_{aq} = \frac{RT}{\log\left(e\right)}\mathrm{p}K_{\mathrm{a}} \tag{5.7}$$

where R is the ideal gas constant and T is the temperature. Experimental pK_{a} values are available from a variety of sources.[341–344] The proton solvation free energy is -265.87 $\pm$ 0.07 kcal/mol, determined by Tissandier *et al.* for the 1M gas phase to 1M solution phase standard state.[227,228] Solvation free energies for many neutral molecules can be found in the compilations referred to above.

Standard state

Equation 5.6 is only valid when all quantities are at the same standard state. Experimental and calculated GPB are reported at a 1 atm standard state, while pK_{a} values are given at the 1 M standard state. Experimental and calculated solvation free energies refer to a 1 M ideal gas entering a 1 M ideal solution. The 1 M standard state and 1

atm standard state are related by $RT\ln\left(\frac{P^\circ(1M)}{P^\circ(1atm)}\right)$, where P° is the vapor pressure at standard state.[339] At 298.15 K to change a 1 atm standard state to a 1 M standard state, 1.89 kcal/mol must be added for the thermodynamic cycle used here.

Electronic structure calculations are microscopic quantities, while experimental values are macroscopic. When a molecule has more than one indistinguishable microscopic protonation state, there will be a difference between microscopic and macroscopic GPB. The conversion between microscopic and macroscopic GPB values is

$$\Delta G^{micro} = \Delta G^{macro} + RT\ln\left(\frac{m}{n}\right) \tag{5.8}$$

where m is the number of indistinguishable microscopic protonation states of A^- and n is the number of indistinguishable microscopic protonation states of HA in Eq. 5.5, R is the ideal gas constant, and T is the temperature of interest as discussed in previous work.[139]

5.2.3 Results and Discussion

Phosphoric acid proton affinity and gas-phase basicity

Table 5.11: Calculated and experimental proton affinities (PA) and gas-phase basicities (GPB) of water, phosphoric acid, dihydrogen phosphate, and hydrogen phosphate.

PA		W1	CBS-QB3	G3B3	QCRNA	QCRNA[a]	Experiment[b]
	H_2O	390.7	392.0	391.5	390.3	392.2	390.3 ± 0.2
	H_3PO_4	328.9	327.9	328.3	328.2	330.0	330.5 ± 5.0
	$H_2PO_4^{1-}$	459.3	458.7	458.6	457.8	459.6	
	HPO_4^{2-}	578.7	580.9	580.9	579.5	581.3	
GPB							
	H_2O	384.1	385.4	384.9	383.8	386.2	383.7 ± 0.2
	H_3PO_4	321.6	320.7	321.0	321.0	323.4	323.2 ± 4.9[c]
	$H_2PO_4^{1-}$	451.6	451.1	451.3	450.0	452.4	
	HPO_4^{2-}	573.2	575.4	572.8	574.1	576.4	

[a] O-H Bond Correction, Ref 139, 140 [b] Ref 121 and 345 [c] Adjusted from macroscopic to microscopic value

The experimental values, as reported by NIST,[121] for the PA and GPB of H_3PO_4 are

330.5 ± 5.0 kcal/mol and 323.0 ± 4.9 kcal/mol, respectively, based on work by Morris *et al.*[345] H_3PO_4 has 4 indistinguishable microscopic protonation states and its conjugate base, $H_2PO_4^-$, has 6 indistinguishable microscopic protonation states. Application of Eq. 5.8 yields a microscopic GPB of 323.2 kcal/mol. No PA or GPB experimental values were found for $H_2PO_4^{1-}$ or HPO_4^{2-}. Table 5.11 summarizes experimental values and the W1 calculations for phosphoric acid and its deprotonation states along with water; also included are calculated values using CBS-QB3,[123,124] G3B3,[125] and QCRNA[97,129] with and without a O-H bond correction.[139]

Of the multi-level methods and QCRNA protocol, W1 performs the best for water and phosphoric acid both for the PA and the GPB. QCRNA using the bond order correction is more accurate than W1 for phosphoric acid, but is 1.9 and 2.5 kcal/mol for the PA and GPB for water, respectively. For the determination of the phosphoric acid solvation free energies, the W1 GPB values are used.

Phosphoric acid solvation free energy

Though experimental solvation data is available for a variety of molecules, reliable experimental data on the solvation of phosphoric acid and its anions is sparse; only two unique literature sources were found and are presented in Table 5.12.

Table 5.12: Experimental literature values for the free energy of solvation (kcal/mol) for phosphoric acid species.

Species	George[a]	Warshel[b]
H_3PO_4	-26	-12[c]
$H_2PO_4^{1-}$	-76	-68
HPO_4^{2-}	-299	-245
PO_4^{3-}	-637	-536

[a] Ref 346. Values are heats of solvation, $\triangle H_{solv}$. [b] Ref 337. See text for details. [c] Calculated using the Iterative Langevin Dipole solvation model.

George *et al.*[346] reports heats of solvation, ($\triangle H_{solv}$), obtained using the Kapustinskii equation to determine the lattice energy and experimental heats of solution. The Kapustinskii equation required the estimation of the cation and anion radii, which were estimated since there is insufficient data to calculate them directly. This value has been

used as a solvation free energy.[108]

Warshel *et al.*[337] reports experimental values only for the three anions, determined from a thermodynamic cycle similar to Equation 5.6. The GPB was obtained by adding -7.5 kcal/mol to the PA taken from a MP2/6-31++G**//HF/6-31G* calculation. The solvation free energy of H_3PO_4 was calculated using the iterative langevin dipoles (ILD) solvation model, presented in the same paper. The solvation free energy of the proton was take as -259.5 kcal/mol. The pK_a values used in the cycle were not reported.

Since the George values are not free energies, the lattice energy was approximated, and there is not an strictly experimental way to determine the solvation entropy of phosphoric acid or its anions, these values cannot be used to derive solvation free energies directly. It is expected that solvation free energies will be less negative than heats of solvation for a polar solute in a polar solvent; therefore, the value for H_3PO_4 can be considered an lower bound for the solvation free energy. The Warshel values were derived using an old value of the proton solvation free energy and a solvation free energy of phosphoric acid calculated using a model that has questionable accuracy for phosphate compounds. Further, the approximation that the difference between PA and GPB is -7.5 kcal/mol seems reasonable when compared to the W1 calculations of H_3PO_4 and $H_2PO_4^{1-}$, whose differences are -7.3 and -7.7 kcal/mol, respectively, but does not agree with the -5.5 kcal/mol difference for HPO_4^{2-}. These approximations and calculations of phosphoric acid solvation reflect the most accurate work at the time, but now these values should only be considered estimates and not experimental values.

Qualitatively, given that phosphoric acid has three hydrogen bond donors, one hydrogen bond acceptor, and a highly polarizable third-row element at the center, it is reasonable to believe that the solvation free energy will be more negative than any neutral organic solutes, which range approximately from 4 kcal/mol (fluoronated alkanes) to -14 kcal/mol (9-methyladenine).[338] In an analysis by Evleth *et al.*[347] the solvation free energy of H_2SO_4 was estimated at -25 kcal/mol, which, along with the heat of solvation from George, further supports a larger negative solvation free energy for phosphoric acid compared to most neutral compounds.

To determine the most reliable solvation free energies possible, an initial benchmark of current implicit solvation models was performed. Table 5.13 summarizes the solvation calculations for the gas-phase and condensed phase optimized geometries, detailed in

Table 5.13: Solvation free energies (kcal/mol) based on gas-phase and solution phase geometry optimizations using various implicit solvation models.

Gas-phase	CPCM	IEFPCM	SCIPCM	COSMO	SM5.43	SM6
H_3PO_4	-21.0	-20.8	-11.1	-11.5	-18.1	-14.9
$H_2PO_4^{1-}$	-69.6	-69.4	-57.1	-67.0	-74.8	-77.6
HPO_4^{2-}	-235.9	-235.8	-207.5	-237.6	-252.9	-265.0
PO_4^{3-}	-526.9	-526.8	-462.0	-530.4	-559.0	-585.9
Solution phase						
H_3PO_4	-22.1	-21.9	-11.3	-12.2	-19.5	-16.2
$H_2PO_4^{1-}$	-72.2	-72.1	-57.8	-68.8	-76.7	-79.6
HPO_4^{2-}	-238.8	-238.7	-208.3	-239.9	-253.8	-266.2
PO_4^{3-}	-528.3	-528.1	-462.8	-531.8	-559.0	-586.2

Section 5.2.2. Using gas-phase optimized geometries or solution optimized geometries made at most 2.9 kcal/mol difference and on average 1.4 kcal/mol, and the general trends for both optimizations are the same. The CPCM and IEFPCM values are similar for all structures, therefore only CPCM will be reported from now on. The SCIPCM values are significantly more positive than all other models. This may be due to its solvation surface response. PCM methods and SM5.43R give the most reasonable values for the solvation free energy of H_3PO_4 based on the George heat of solvation and the arguments given above.

Table 5.14: Experimental pK_a values (Ref 342) and W1 calculated gas-phase basicities (See Table 5.11) for phosphoric acid, dihydrogen phosphate, and hydrogen phosphate for different standard states.

	pK_a		Gas-phase Basicity		
	micro	macro	micro,1atm	macro,1atm	macro,1M
$H_3PO_4^{1-}$	1.94	2.21	321.6	321.4	323.3
$H_2PO_4^{2-}$	7.38	7.21	451.6	451.9	453.8
HPO_4^{3-}	13.24	12.67	573.2	572.4	574.3

Though absolute experimental solvation free energies are not available, using Equation 5.6 differences between solvation energies can be derived

$$\begin{aligned} \Delta\Delta G^{\circ}_{solv}\left(A^-, HA\right) &= \Delta G^{\circ}_{solv}\left(A^-\right) - \Delta G^{\circ}_{solv}\left(HA\right) \\ &= \frac{RT}{log(e)} pK_a - \Delta G^{\circ}_{gas} - \Delta G^{\circ}_{solv}\left(H^+\right) \end{aligned} \tag{5.9}$$

Using the W1 GPB values adjusted for standard state, the experimental pK_a values[342] (shown in Table 5.14), and the solvation free energy of the proton[227] with Equation 5.9, the three solvation free energy differences are derived and presented in Table 5.15, denoted as predicted values. Given the error in W1 is < 1 kcal/mol, the error in the proton is 0.1 kcal/mol, and the error in pK_a values is estimated at < 0.3 kcal/mol given the range in values from Perrin *et al.*,[344] these differences are assigned an error of $\pm$ 1.4 kcal/mol.

Table 5.15: Solvation free energies differences (kcal/mol) from structures optimized in the gas-phase and solution.

Gas-Phase	CPCM	SCIPCM	COSMO	SM5.43	SM6	Predicted
$H_2PO_4^{1-}$ - H_3PO_4	-48.6	-46.0	-55.6	-56.7	-62.7	-54.5
HPO_4^{2-} - $H_2PO_4^{1-}$	-166.3	-150.4	-170.6	-178.1	-187.4	-178.1
PO_4^{3-} - HPO_4^{2-}	-291.0	-254.5	-292.8	-306.1	-320.9	-291.2
Solution Phase						
$H_2PO_4^{1-}$ - H_3PO_4	-50.1	-46.5	-56.6	-57.2	-63.4	-54.5
HPO_4^{2-} - $H_2PO_4^{1-}$	-166.6	-150.5	-171.1	-177.1	-186.7	-178.1
PO_4^{3-} - HPO_4^{2-}	-289.5	-254.6	-292.0	-305.2	-320.0	-291.2

Table 5.15 also provides the solvation free energy differences of the calculated implicit solvation models. Again, gas-phase optimization compares well to the solution phase structures. While the PCM model seemed to be a good model for H_3PO_4, its relative solvation to the anion is underestimated as well as for the second deprotonation. COSMO, which underestimated the solvation energy of H_3PO_4, has a very reasonable relative solvation for the three deprotonations. The SM5.43 model describes the first two deprotonations, but significantly errs on the last. Overall, while an individual model can describe one or at best two deprotonations, none are able to get the relative solvation within the 1.4 kcal/mol necessary to get a prediction within ± 1 pK_a unit.

Suggested value for ΔG_{aq} for phosphoric acid and its anions

To provide the best estimate for the solvation energies of phosphoric acid and its anions, a combination of all available data should be used. Because each of the solvation energies are connected through Equation 5.9, by knowing the GPB, proton solvation free

energy, and the pK_a, in actuality only the solvation free energy of one species is needed. While the direct experimental determination of the partition coefficient between vapor and water has not yet been done due to its technical difficulty, solvation free energies have been measured for mono-, di-, and triester phosphates from water to $CHCl_3$ and triester phosphates for water to vapor by Wolfenden and Williams.[328] These values are reproduced in Table 5.16.

Table 5.16: Partition coefficients for transfer of phosphorus derivatives from water to nonpolar environments (chloroform and vapor) at 20 °C, ionic strength 0.30. All values reproduced from Reference 328.

	Phosphate Ester	$CHCl_3$	Vapor
tri-	$(PrO)_3PO$	$4.7x10^{3}$	$2.8x10^{-5}$
	$(EtO)_3PO$	$1.9x10^{2}$	$1.5x10^{-6}$
	$(MeO)_3PO$	5.8	$3.0x10^{-7}$
di-	$(PrO)_2PO(OH)$	$9.0x10^{-2}$	
	$(EtO)_2PO(OH)$	$2.7x10^{-3}$	
	$(MeO)_2PO(OH)$	$1.5x10^{-4}$	
mono-	$(PrO)PO(OH)_2$	$2.5x10^{-6}$	

Based on the n-propyl phosphate ester, the replacement of a propyl group to a hydroxyl group lowers the solvation into chloroform by 6.1 and 6.3 kcal/mol. Similarly, for an ethyl replacement this is 6.5 kcal/mol and for methyl replacement is 6.1 kcal/mol. If we assume that this change is similar for transfer into the vapor from water, this gives estimates of for the solvation free energy (vapor to water) of -24.8 (propyl), -27.3 (ethyl), and -27.2 (methyl) kcal/mol for phosphoric acid.

This assumes that the alkane substituents have the same effect in chloroform as they would in vapor, but this is obviously not true as one would expect the alkane substituents to favor a transfer from vapor to chloroform. To correct for the chloroform phase, we can use the estimate from Wolfenden and Lewis that a -CH2- group disfavors the gas-phase over chloroform by 0.8 kcal/mol.[348] This raises our estimates to -17.6 (propyl), -22.5 (ethyl), and -24.8 (methyl) kcal/mol.

All of these fall above the lower bound of George's heat of solvation, -26 kcal/mol,

and is consistent, though certainly high for the propyl estimation, with Evleth's estimation of H_2SO_4 solvation free energy. These estimates also trend lower as smaller alkane esters are used, making it likely that the phosphoric acid limit is likely closer to methyl value than the propyl. This leads to the best possible estimate for the solvation free energy of phosphoric acid around -25.0 kcal/mol. Given that phosphoric acid presents 4 hydrogen bonding donor/acceptor sites, it is reasonable to take the entropy contribution to be small. Using Equation 5.6, this gives us solvation free energies of -25.0 kcal/mol for H_3PO_4, -79.5 kcal/mol for $H_2PO_4^-$, -257.6 kcal/mol for HPO_4^{2-}, and -548.8 kcal/mol for PO_4^{3-}.

5.2.4 Conclusion

Understanding the solvation of phosphoric acid and its anions is an important benchmark both for implicit solvation model development as well as a useful experimental benchmark. We suggest using solvation free energies of -25.0 kcal/mol for H_3PO_4, -79.5 kcal/mol for $H_2PO_4^-$, -257.6 kcal/mol for HPO_4^{2-}, and -548.8 kcal/mol for PO_4^{3-} in future work. These values are consistent with known pK_a values and the calculated gas-phase basicities. Provided are current predictions of these solvation energies based on various implicit solvation models. None of the models tested here are able to reproduce all the solvation values for the three deprotonations of phosphoric acid.

5.2.5 Acknowledgment

For this work, special thanks is given to Dr. R. Wolfenden for useful discussion on using alkylated phosphate esters to estimate the affect of nonpolar groups on solvation.

References

[1] I. Vidovic, S. Nottrott, K. Hartmuth, R. Luhrmann, and R. Ficner, Mol. Cell **6**, 1331 (2000).

[2] N. Toor, K. S. Keating, S. D. Taylor, and A. M. Pyle, Science **320**, 77 (2008). URL `http://dx.doi.org/10.1126/science.1153803`.

[3] E. Uhlmann and A. Peyman, Chem. Rev. **90**, 543 (1990).

[4] N. Sarver, E. M. Cantin, P. S. Chang, J. A. Zaia, P. A. Ladne, D. A. Stephens, and J. J. Rossi, Science **247**, 1222 (1990).

[5] S. Altman, Proc. Natl. Acad. Sci. USA **90**, 10898 (1993).

[6] L. N. Buryyanovskii and A. D. Shved, Biopolim. Kletka **12**, 20 (1996).

[7] N. Usman, L. Beigelman, and J. A. McSwiggen, Curr. Opin. Struct. Biol. **1**, 527 (1996).

[8] A. R. Muotri, L. da Veiga Pereira, L. dos Reis Vasques, and C. F. M. Menck, Gene **237**, 303 (1999).

[9] J. T. Holmlund, Curr. Opin. Mol. Ther. **1**, 372 (1999).

[10] M. D. Hughes, M. Hussain, Q. Nawaz, P. Sayyed, and S. Akhtar, DDT **6**, 313 (2001).

[11] S. P. Zinnen, K. Domenico, M. Wilson, B. Dickinson, A. Beaudry, V. Mokler, A. T. Daniher, A. Burgin, and L. Beigelman, RNA **8**, 214 (2002).

[12] D. Maniotis, M. Wood, and L. Phylactou, Neurosci. Lett. **329**, 81 (2002).

[13] C. D. Novina, M. F. Murray, D. M. Dykxhoorn, P. J. Beresford, J. Riess, S.-K. Lee, R. G. Collman, J. Lieberman, P. Shankar, and P. A. Sharp, Nat. Med. **8**, 681 (2002).

[14] B. Sriram, D. Thakral, and S. K. Panda, Virology **312**, 350 (2003).

[15] S. M. L. Raj and F. Liu, Gene **313**, 59 (2003).

[16] A. Tekos, C. Stathopoulos, D. Tsambaos, and D. Drainas, Curr. Med. Chem. **11**, 2979 (2004).

[17] L. M. Alvarez-Salas, M. L. Benítez-Hess, and J. A. DiPaolo, Antivir. Ther. **8**, 265 (2003).

[18] M. Rubenstein, R. Tsui, and P. Guinan, Drugs of the Future **29**, 893 (2004).

[19] T. R. Cech, Curr. Opin. Struct. Biol. **2**, 605 (1992).

[20] R. R. Breaker, Chem. Rev. **97**, 371 (1997).

[21] P. T. Sekella, D. Rueda, and N. G. Walter, RNA **8**, 1242 (2002).

[22] E. Puerta-Fernández, C. Romero-López, A. Barroso-delJesus, and A. Berzal-Herranz, FEMS Microbiol. Rev. **27**, 75 (2003).

[23] A. Lescoute and E. Westhof, Chem. Biol. **12**, 10 (2005).

[24] G. A. Soukup and R. R. Breaker, Trends Biotechnol. **17**, 469 (1999).

[25] G. A. Soukup and R. R. Breaker, Proc. Natl. Acad. Sci. **96**, 3584 (1999).

[26] G. A. Soukup, G. A. M. Emilsson, and R. R. Breaker, J. Mol. Biol. **298**, 623 (2000).

[27] G. A. Soukup and R. R. Breaker, Curr. Opin. Struct. Biol. **10**, 318 (2000).

[28] M. Koizumi, G. A. Soukup, J. N. Q. Kerr, and R. R. Breaker, Nature Struct. Biol. **6**, 1062 (1999).

[29] T. Kuwabara, M. Warashina, and K. Taira, Trends Biotech. **18**, 462 (2000).

[30] S. Seetharaman, M. Zivarts, N. Sudarsan, and R. R. Breaker, Nature Biotech. **19**, 336 (2001).

[31] N. K. Vaish, F. Dong, L. Andrews, R. E. Schweppe, N. G. Ahn, L. Blatt, and S. D. Seiwert, Nature Biotech. **20**, 810 (2002).

[32] D. H. Burke, N. D. S. Ozerova, and M. Nilsen-Hamilton, Biochemistry **41**, 6588 (2002).

[33] J. C. Achenbach, R. Nutiu, and Y. Li, Anal. Chim. Acta **534**, 41 (2005).

[34] L. J. Bergeron and J.-P. Perreault, Nucleic Acids Res. **33**, 1240 (2005).

[35] M. Famulok, Curr. Opin. Struct. Biol. **9**, 324 (1999).

[36] K. A. Marshall and A. D. Ellington, Nature Struct. Biol. **6**, 992 (1999).

[37] T. Kuwabaraa, M. Warashinab, and K. Taira, Curr. Opin. Chem. Biol. **4**, 669 (2000).

[38] R. R. Breaker, Curr. Opin. Biotechnol. **13**, 31 (2002).

[39] J. M. Yeakley, J.-B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M. S. Chee, and X.-D. Fu, Nat. Biotechnol. **20**, 353 (2002).

[40] S. K. Silverman, RNA **9**, 377 (2003).

[41] C. A. Collins and C. Guthrie, Nature Struct. Biol. **7**, 850 (2000).

[42] S. Valadkhan and J. L. Manley, Nature **413**, 701 (2001).

[43] S. Valadkhan, Science **307**, 863 (2005).

[44] A. Ke, F. Ding, J. D. Batchelor, and J. A. Doudna, Structure **15**, 281 (2007).

[45] J. M. Burke, Nature Struct. Biol. **8**, 382 (2001).

[46] E. Westhof, J. Mol. Recog. **20**, 1 (2006).

[47] D. M. Lilley and F. Eckstein, editors, *Ribozymes and RNA Catalysis*, RSC Biomolecular Series (RSC Publishing, Cambridge, 2008).

[48] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics* (John Wiley and Sons, New York, 1977).

[49] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, 1st ed. (Dover Publications, Inc., New York, 1996).

[50] C. C. J. Roothaan, Rev. Mod. Phys. **23**, 69 (1951).

[51] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd ed. (John Wiley & Sons, Chichester, England, 2004).

[52] M. J. S. Dewar, E. Zoebisch, E. F. Healy, and J. J. P. Stewart, J. Am. Chem. Soc. **107**, 3902 (1985).

[53] J. J. P. Stewart, J. Comput. Chem. **10**, 209 (1989).

[54] J. J. P. Stewart, J. Mol. Model. **13**, 1173 (2007).

[55] X. Lopez and D. M. York, Theor. Chem. Acc. **109**, 149 (2003).

[56] T. J. Giese, E. C. Sherer, C. J. Cramer, and D. M. York, J. Chem. Theory Comput. **1**, 1275 (2005).

[57] K. Nam, Q. Cui, J. Gao, and D. M. York, J. Chem. Theory Comput. **3**, 486 (2007).

[58] J. Khandogin, A. Hu, and D. M. York, J. Comput. Chem. **21**, 1562 (2000).

[59] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[60] R. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

[61] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, 2nd ed. ed. (WILY-VCH, Weinhein, Germany, 2001).

[62] A. D. Becke, Phys. Rev. A. **38**, 3098 (1988).

[63] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[64] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B. **37**, 785 (1988).

[65] L. A. Curtiss, K. Raghavachari, and G. W. Trucks, J. Chem. Phys. **94**, 7221 (1991).

[66] T.-S. Lee, D. M. York, and W. Yang, J. Chem. Phys. **105**, 2744 (1996).

[67] M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1987).

[68] L. Greengard, *The rapid evaluation of potential fields in particle systems* (The MIT Press, Cambridge, MA, 1988).

[69] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).

[70] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[71] S. W. Rick, J. Chem. Phys. **120**, 6085 (2004).

[72] J. Zielkiewicz, J. Chem. Phys. **123**, 104501 (2005).

[73] S. E. Wong, K. Bernacki, and M. Jacobson, J. Phys. Chem. B **109**, 5249 (2005).

[74] B. Roux and T. Simonson, Biophys. Chem. **78**, 1 (1999).

[75] C. J. Cramer and D. G. Truhlar, Chem. Rev. **99**, 2161 (1999).

[76] M. Orozco and L. F. Javier, Chem. Rev. **100**, 4187 (2000).

[77] J. Tomasi and M. Persico, Chem. Rev. **94**, 2027 (1994).

[78] K. Sharp and B. Honig, J. Phys. Chem. **94**, 7684 (1990).

[79] V. Dillet, D. Rinaldi, and J.-L. Rivail, J. Phys. Chem. **98**, 5034 (1994).

[80] S. Miertuš, E. Scrocco, and J. Tomasi, Chem. Phys. **55**, 117 (1981).

[81] M. Cossi, V. Barone, R. Cammi, and J. Tomasi, Chem. Phys. Lett. **255**, 327 (1996).

[82] A. Klamt and G. Schüürmann, J. Chem. Soc. Perkin Trans. 2 **2**, 799 (1993).

[83] T. N. Truong and E. V. Stefanovich, Chem. Phys. Lett. **240**, 253 (1995).

[84] D. M. York and M. Karplus, J. Phys. Chem. A **103**, 11060 (1999).

[85] D. Giesen, C. Cramer, and D. Truhlar, J. Phys. Chem. **99**, 7137 (1995).

[86] C. J. Cramer and D. G. Truhlar, J. Comput.-Aided Mol. Des. **6**, 629 (1992).

[87] C. P. Kelly, C. J. Cramer, and D. G. Truhlar, J. Chem. Theory Comput. **1**, 1177 (2005).

[88] T. J. Giese, B. A. Gregersen, Y. Liu, K. Nam, E. Mayaan, A. Moser, K. Range, O. Nieto Faza, C. Silva Lopez, A. Rodriguez de Lera, G. Schaftenaar, X. Lopez, T. Lee, G. Karypis, and D. M. York, J. Mol. Graph. Model. **25**, 423 (2006).

[89] B. A. Gregersen and D. M. York, J. Phys. Chem. B **109**, 536 (2005).

[90] B. A. Gregersen and D. M. York, J. Comput. Chem. **27**, 103 (2006).

[91] R. B. Silverman, *The Organic Chemistry of Enzyme-Catalyzed Reactions* (Academic Press, San Diego, CA, 2000).

[92] Y. Alexeev, T. Windus, C.-G. Zhan, and D. Dixon, Int. J. Quantum Chem. **102**, 775 (2005).

[93] G. I. Almerindo, D. W. Tondo, and J. R. Pliego Jr., J. Phys. Chem. A **108**, 166 (2004).

[94] Y. Fu, L. Liu, R.-Q. Li, R. Liu, and Q.-X. Guo, J. Am. Chem. Soc. **126**, 814 (2004).

[95] P. Hudáky and A. Perczel, J. Phys. Chem. A **108**, 6195 (2004).

[96] A. M. Magill, K. J. Cavell, and B. F. Yates, J. Am. Chem. Soc. **126**, 8717 (2004).

[97] K. Range, M. J. McGrath, X. Lopez, and D. M. York, J. Am. Chem. Soc. **126**, 1654 (2004).

[98] Y. H. Jang, W. A. Goddard III, K. T. Noyes, L. C. Sowers, S. Hwang, and D. S. Chung, J. Phys. Chem. B **107**, 344 (2003).

[99] A. Klamt, F. Eckert, M. Diedenhofen, and M. E. Beck, J. Phys. Chem. A **107**, 9380 (2003).

[100] K. R. Adam, J. Phys. Chem. A **106**, 11963 (2002).

[101] D. M. Chipman, J. Phys. Chem. A **106**, 7413 (2002).

[102] J. J. Klicić, R. A. Friesner, S.-Y. Liu, and W. C. Guida, J. Phys. Chem. A **106**, 1327 (2002).

[103] X. Lopez, M. Schaefer, A. Dejaegere, and M. Karplus, J. Am. Chem. Soc. **124**, 5010 (2002).

[104] J. R. Pliego, Jr. and J. M. Riveros, J. Phys. Chem. A **106**, 7434 (2002).

[105] M. D. Liptak and G. C. Shields, J. Am. Chem. Soc. **123**, 7314 (2001).

[106] M. D. Liptak and G. C. Shields, Int. J. Quantum Chem. **85**, 727 (2001).

[107] I.-J. Chen and A. D. MacKerell Jr, Theor. Chem. Acc. **103**, 483 (2000).

[108] P. D. Lyne and M. Karplus, J. Am. Chem. Soc. **122**, 166 (2000).

[109] C. O. Silva, E. C. da Silva, and M. A. C. Nascimento, J. Phys. Chem. A **104**, 2402 (2000).

[110] J. E. Yazal, F. G. Prendergast, D. E. Shaw, and Y.-P. Pang, J. Am. Chem. Soc. **122**, 11411 (2000).

[111] M. Peräkylä, Phys. Chem. Chem. Phys. **1**, 5643 (1999).

[112] C. O. da Silva, E. C. da Silva, and M. A. C. Nascimento, J. Phys. Chem. A **103**, 11194 (1999).

[113] G. Schüürmann, M. Cossi, V. Barone, and J. Tomasi, J. Phys. Chem. A **102**, 6706 (1998).

[114] W. H. Richardson, C. Peng, D. Bashford, L. Noodleman, and D. A. Case, Int. J. Quantum Chem. **61**, 207 (1997).

[115] H. Li, A. D.Robertson, and J. H.Jensen, Proteins **55**, 689 (2004).

[116] A. D. McNaught and A. Wilkinson, *Compendium of Chemical Terminology: IUPAC Recommendations*, 2nd ed. (Blackwell Science, Inc., Oxford, 1997). `http://www.iupac.org/publications/compendium/index.html`.

[117] K. Nam, J. Gao, and D. M. York, *Multiscale Simulation Methods for Nanomaterials*, edited by R. B. Ross and M. Sanat, 201–218 (Wiley, 2008).

[118] R. B. Martin, Acc. Chem. Res. **18**, 32 (1985).

[119] J. J. Dannenberg and M. Tomasz, J. Am. Chem. Soc. **122**, 2062 (2000).

[120] A. Moser, R. Guza, N. Tretyakova, and D. M. York, Theor. Chem. Acc. **122**, 179 (2009).

[121] P. Linstrom and W. Mallard, editors, *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (National Institute of Standards and Technology, Gaithersburg MD, 20899, 2003). (`http://webbook.nist.gov`).

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

Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, Gaussian 03, Revision C.02 (2004). Gaussian, Inc., Wallingford, CT.

[123] J. A. Montgomery, Jr., M. J. Frisch, J. W. Ochterski, and G. A. Petersson, J. Chem. Phys. **110**, 2822 (1999).

[124] J. A. Montgomery, Jr., M. J. Frisch, J. W. Ochterski, and G. A. Petersson, J. Chem. Phys. **112**, 6532 (2000).

[125] A. G. Baboul, L. A. Curtiss, P. C. Redfern, and K. Raghavachari, J. Chem. Phys. **110**, 7650 (1999).

[126] E. K. Pokon, M. D. Liptak, S. Feldgus, and G. C. Shields, J. Phys. Chem. A **105**, 10483 (2001).

[127] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, and J. A. Pople, J. Chem. Phys. **109**, 7764 (1998).

[128] C. S. López, O. N. Faza, B. A. Gregersen, X. Lopez, A. R. de Lera, and D. M. York, Chem. Phys. Chem. **5**, 1045 (2004).

[129] E. Mayaan, K. Range, and D. M. York, J. Biol. Inorg. Chem. **9**, 807 (2004).

[130] C. S. López, O. N. Faza, A. R. de Lera, and D. M. York, Chem. Eur. J. **11**, 2081 (2005).

[131] C. Adamo and G. E. Scuseria, J. Chem. Phys. **111**, 2889 (1999).

[132] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[133] A. D. Becke, J. Chem. Phys. **104**, 1040 (1996).

[134] Y. Zhao and D. G. Truhlar, J. Chem. Theory Comput. **1**, 415 (2005).

[135] J. W. Ochterski, Vibrational Analysis in *Gaussian*, `http://gaussian.com/g_whitepap/vib.htm` (accessed March 2005) (1999).

[136] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd ed. (John Wiley & Sons, Chichester, England, 2002).

[137] D. A. McQuarrie, *Statistical Mechanics* (University Science Books, Mill Valley, CA, 1973).

[138] H. Lioe, R. A. O'Hair, S. Gronert, A. Austin, and G. E. Reid, Int. J. Mass Spectrom. **267**, 220 (2007).

[139] K. Range, D. Riccardi, Q. Cui, M. Elstner, and D. M. York, Phys. Chem. Chem. Phys. **7**, 3070 (2005).

[140] K. Range, C. S. López, A. Moser, and D. M. York, J. Phys. Chem. A **110**, 791 (2006).

[141] F. R. Tortonda and J. L. Pascual-Ahuir and E. Silla and I. Tuñón, Chem. Phys. Lett. **260**, 21 (1996).

[142] G. Yang, Y. Zu, C. Liu, Y. Fu, and L. Zhou, J. Phys. Chem. B **112**, 7104 (2008).

[143] C. M. Jones, M. Bernier, E. Carson, K. E. Colyer, R. Metz, A. Pawlow, E. D. Wischow, I. Webb, E. J. Andriole, and J. C. Poutsma, Int. J. Mass Spectrom. **267**, 54 (2007).

[144] Borislav Kovačevic and Marko Rožman and Leo Klasinc and Dunja Srzić and Zvonimir B. Maksić and Manuel Yáñez, J. Phys. Chem. A **109**, 8329 (2005).

[145] A. F. Kuntz, A. W. Boynton, G. A. David, K. E. Colyer, and J. C. Poutsma, J. Am. Soc. Mass Spectrom. 72–81 (2002).

[146] D. Voet, J. Voet, and C. Pratt, *Fundamentals of Biochemistry* (John Wiley & Sons, Inc., New York, New York, 1999).

[147] E. Hunter and S. Lias, J. Phys. Chem. Ref. Data **27**, 413 (1998).

[148] A. D. McNaught and A. Wilkinson, editors, *IUPAC Compendium of Chemical Terminology (the Gold Book)*, 2nd ed. (Blackwell Scientific Publications, Oxford, 1997).

[149] J. Lee, Int. J. Mass Spectrom. **240**, 261 (2005).

[150] L. D. Donna, A. Napoli, and G. Sindona, J. Am. Soc. Mass. Spectrom. **15**, 1080 (2004).

[151] V. A. Bloomfield, D. M. Crothers, and I. Tinoco, Jr., *Nucleic Acids: Structures, Properties, and Functions* (University Science Books, Sausalito, CA, 2000).

[152] F. Greco, A. Liguori, G. Sindona, and N. Uccella, J. Am. Chem. Soc. **112**, 9092 (1990).

[153] J. K. Wolken and F. Tureček, J. Am. Soc. Mass Spectrom. **11**, 1065 (2000).

[154] R. Wu and T. B. McMahon, J. Am. Chem. Soc. **129**, 569 (2007).

[155] Nino Russo and Marrirosa Toscano and Andrè Grand and Franck Jolibois, J. Comput. Chem. **19**, 989 (1998).

[156] Y. Podolyan, L. Gorb, and J. Leszczynski, J. Phys. Chem. A **104**, 7346 (2000).

[157] Y. Li and R. R. Breaker, J. Am. Chem. Soc. **121**, 5364 (1999).

[158] S. Acharya, A. Földesi, and J. Chattopadhyaya, J. Org. Chem. **68**, 1906 (2003).

[159] C. Altona and M. Sundaralingam, J. Am. Chem. Soc. **94**, 8205 (1972).

[160] S. Arnott and D. Hukins, Biochem. J. **130**, 453 (1972).

[161] S. Arnott and D. W. L. Hukins, Biochem. Biophys. Res. Commun. **47**, 1504 (1972).

[162] W. Saenger, *Principles of nucleic acid structure* (Springer-Verlag, New York, 1984).

[163] P. C. Bevilacqua, *Ribozymes and RNA Catalysis*, chap. Proton Transfer in Ribozyme Catalysis, 11–36 (RSC Publishing, 2008).

[164] X. Lopez, D. M. York, A. Dejaegere, and M. Karplus, Int. J. Quantum Chem. **86**, 10 (2002).

[165] X. Lopez, A. Dejaegere, F. Leclerc, D. M. York, and M. Karplus, J. Phys. Chem. B **110**, 11525 (2006).

[166] A. E. A. Hassan, J. Sheng, J. Jiang, W. Zhang, and Z. Huang, Org. Lett. **12**, 2503 (2009).

[167] A. C. Hengge, Adv. Phys. Org. Chem. **40**, 49 (2005).

[168] M. Bianciotto, J.-C. Barthelat, and A. Vigroux, J. Am. Chem. Soc. **124**, 7573 (2002).

[169] R. Wolfenden, C. Ridgway, and G. Young, J. Am. Chem. Soc. **120**, 833 (1998).

[170] J. Khandogin, B. A. Gregersen, W. Thiel, and D. M. York, J. Phys. Chem. B **109**, 9799 (2005).

[171] K. Nam, J. Gao, and D. M. York, J. Chem. Theory Comput. **1**, 2 (2005).

[172] K. N. Allen and D. Dunaway-Mariano, Trends Biochem. Sci. **29**, 495 (2004).

[173] K. C. K. Swamy and N. S. Kumar, Curr. Sci. **85**, 1256 (2003).

[174] P. A. Frey and R. D. Sammons, Science **228**, 541 (1985).

[175] D. Herschlag, J. A. Piccirilli, and T. R. Cech, Biochemistry **30**, 4844 (1991).

[176] B. A. Gregersen, J. Khandogin, W. Thiel, and D. M. York, J. Phys. Chem. B **109**, 9810 (2005).

[177] Y. Liu, B. A. Gregersen, X. Lopez, and D. M. York, J. Phys. Chem. B **109**, 19987 (2005).

[178] J.-P. Jost and J. Hofsteenge, Proc. Natl. Acad. Sci. USA **89**, 9699 (1992).

[179] M. R. Holman, T. Ito, and S. E. Rokita, J. Am. Chem. Soc. **129**, 6 (2007).

[180] M. Ehrlich, M. A.Gama-Sosa, L.-H. Huang, R. M. Midgett, K. C. Kuo, R. A.McCune, and C. Gehrke, Nucleic Acids Res. **10**, 2709 (1982).

[181] E. N. Gal-Yam, Y. Saito, G. Egger, and P. A. Jones, Annu. Rev. Med. **59**, 267 (2008).

[182] A. Bird, Nature **447**, 396 (2007).

[183] G. A. Romanov, E. N. Zhavoronkova, S. V. Savel'ev, and B. F. Vanyushin, Neurosci. Behav. Physiol. **16**, 285 (1983).

[184] D. Yu, J. A. Berlin, T. M. Penning, and J. Field, Chem. Res. Toxical. **15**, 832 (2002).

[185] F. H. Hausheer, S. N. Rao, M. P. Gamcsik, P. A. Kollman, O. M. Colvin, J. D. Saxe, B. D. Nelkin, I. J. McLennan, G. Barnett, and S. B. Baylin, Carcinogenesis **10**, 1131 (1989).

[186] A. P. Bird, Nature **321**, 209 (1986).

[187] A. D. Riggs, Adv. Cancer Res. **40**, 1 (1983).

[188] M. F. Denissenko, J. X. Chen, M.-S. Tang, and G. P. Pfeifer, Proc. Natl. Acad. Sci. USA **94**, 3893 (1997).

[189] T. M. Hernandez-Boussard and P. Hainaut, Environ. Health Prospect. **106**, 385 (1998).

[190] B. Matter, G. Wang, R. Jones, and N. Tretyakova, Chem. Res. Toxicol. **17**, 731 (2004).

[191] M. F. Denissenko, A. Pao, M. shong Tang, and G. P. Pfeifer, Science **274**, 430 (1996).

[192] D. Hoffmann and I. Hoffmann, J. Toxicol. Environ. Health **50**, 307 (1997).

[193] S. S. Hecht, J. Natl. Cancer Inst. **91**, 1194 (1999).

[194] S. P. Hussain and C. C. Harris, Mutat. Res. **428**, 23 (1999).

[195] K. Sendowski and M. F. Rajewsky, Mutat. Res. **250**, 153 (1991).

[196] B. H. Mathison, B. Said, and R. C. Shank, Carcinogenesis **14**, 323 (1993).

[197] J. X. Chen, Y. Zheng, M. West, and M. shong Tang, Cancer Res. **58**, 2070 (1998).

[198] D. J. Weisenberger and L. J. Romano, J. Biol. Chem. **274**, 23948 (1999).

[199] A. Das, K. S. Tang, S. Gopalakrishnan, M. J. Waring, and M. Tomasz, Chem. Biol. **6**, 461 (1999).

[200] M. K. Ross, B. H. Mathison, B. Said, and R. C. Shank, Biochem. Biophys. Res. Commun. **254**, 114 (1999).

[201] V.-S. Li, M. Reed, Y. Zheng, H. Kohn, and M. shong Tang, Biochemistry **39**, 2612 (2000).

[202] G. P. Pfeifer, M. Tang, and M. F. Denissenko, Curr. Top. Micro. Biol. Immunol. **249**, 1 (2000).

[203] A. Burdzy, K. T. Noyes, V. Valinluck, and L. C. Sowers, Nucleic Acids Res. **30**, 4068 (2002).

[204] M. Rajesh, G. Wang, R. Jones, and N. Tretyakova, Biochemistry **44**, 2197 (2005).

[205] S. S. Hecht, J. Natl. Cancer Inst. **92**, 782 (2000).

[206] N. Tretyakova, B. Matter, R. Jones, and A. Shallop, Biochemistry **41**, 9535 (2002).

[207] R. Ziegel, A. Shallop, P. Upadhyaya, R. Jones, and N. Tretyakova, Biochemistry **43**, 540 (2004).

[208] N. Zhang, C. Lin, X. Huang, A. Kolbanovskiy, B. E. Hingerty, S. Amin, S. Broyde, N. E. Geacintov, and D. J. Patei, J. Mol. Biol. **346**, 951 (2005).

[209] F. A. Rodríguez, Y. Cai, C. Lin, Y. Tang, A. Kolbanovskiy, S. Amin, D. J. Patel, S. Broyde, and N. E. Geacintov, Nucleic Acids Res. **35**, 1555 (2007).

[210] V. Valinluck, P. Liu, J. I. K. Jr, A. Burdzy, and L. C. Sowers, Nucleic Acids Res. **33**, 3057 (2005).

[211] Y. H. Jang, L. C. Sowers, T. Çağin, and W. A. Goddard III, J. Phys. Chem. A **105**, 274 (2001).

[212] F. Meng, C. Liu, and W. Xu, Chem. Phys. Lett. **373**, 72 (2003).

[213] E. S. Kryachko and M. T. Nguyen, J. Phys. Chem. A **106**, 9319 (2002).

[214] E. C. Sherer, D. M. York, and C. J. Cramer, J. Comput. Chem. **24**, 57 (2003).

[215] J. Šponer, J. Leszczynski, and P. Hobza, J. Mol. Struct. (Theochem) **573**, 43 (2001).

[216] D. Řeha, M. Kabeláč, F. Ryjáček, J. Šponer, J. E. Š amd Marcus Elstner, S. Suhai, and P. Hobza, J. Am. Chem. Soc. **124**, 3366 (2002).

[217] A. Kumar, M. Elstner, and S. Suhai, Int. J. Quantum Chem. **95**, 44 (2003).

[218] I. Dabkowska, P. Jurečka, and P. Hobza, J. Chem. Phys. **122**, 204322 (2005).

[219] QCRNA, *http://theory.chem.umn.edu/Database/QCRNA*.

[220] R. Bauernschmitt and R. Ahlrichs, J. Chem. Phys. **104**, 9047 (1996).

[221] R. Seeger and J. A. Pople, J. Chem. Phys. **66**, 3045 (1977).

[222] A. E. Reed, R. B. Weinstock, and F. Weinhold, J. Chem. Phys. **83**, 735 (1985).

[223] Æleen Frisch and Michael J. Frisch, *Gaussian 98 User's Reference*, 2$^{\text{nd}}$ ed. (Gaussian, Inc., Pittsburgh, PA, 1999).

[224] T. Mineva, N. Russo, and E. Sicilia, J. Comput. Chem. **19**, 290 (1998).

[225] V. Barone and M. Cossi, J. Phys. Chem. A **102**, 1995 (1998).

[226] G. Li and Q. Cui, J. Phys. Chem. B **107**, 14521 (2003).

[227] M. D. Tissandier, K. A. Cowen, W. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, and T. R. Tuttle, Jr., J. Phys. Chem. A **102**, 7787 (1998).

[228] D. M. Camaioni and C. A. Schwerdtfeger, J. Phys. Chem. A **109**, 10795 (2005).

[229] S. Kawahara, A. Kobori, M. Sekine, K. Taira, and T. Uchimaru, J. Phys. Chem. A **105**, 10596 (2001).

[230] A. Asensio, N. Kobko, and J. J. Dannenberg, J. Phys. Chem. A **107**, 6441 (2003).

[231] X. Zhang and C. K. Mathews, J. Biol. Chem. **26**, 7066 (1994).

[232] L. C. Sowers, J. Biomol. Struct. Dyn. **17**, 713 (2000).

[233] I. K. Yanson, A. B. Teplitsky, and L. F. Sukhodub, Biopolymers **18**, 1149 (1979).

[234] J. Šponer, J. Leszczynski, and P. Hobza, Biopolymers **61**, 3 (2002).

[235] J. Šponer, P. Jurečka, and P. Hobza, J. Am. Chem. Soc. **126**, 10142 (2004).

[236] Y. Mo, J. Mol. Model **12**, 665 (2006).

[237] V. Valinluck, W. Wu, P. Liu, J. W. Neidigh, and L. C. Sowers, Chem. Res. Toxicol. **19**, 556 (2006).

[238] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry* (W. H. Freeman and Co., New York, New York, 2002).

[239] J. Norberg and M. Vihinen, J. Mol. Struct. Theochem **546**, 51 (2001).

[240] N. E. Geacintov, H. Yoshia, V. Ibanez, S. A. Jacobs, and R. G. Harvey, Biochem. Biophys. Res. Commun. **122**, 33 (1984).

[241] L. C. Sowers, B. R. Shaw, and W. D. Sedwick, Biochem. Biophys. Res. Commun. **148**, 790 (1987).

[242] I. A. Blair, J. Biol. Chem. **283**, 15545 (2008).

[243] I.-Y. Yang, K. Hashimoto, N. de Wind, I. A. Blair, and M. Moriya, J. Biol. Chem. **284**, 191 (2009).

[244] M. H. G. Medeiros, Chem. Res. Toxicol. **22**, 419 (2009).

[245] D. T. Nair, R. E. Johnson, L. Prakash, S. Prakash, and A. K. Aggarwal, Nat. Struct. Mol. Biol. **13**, 619 (2006).

[246] S. P. Fink, G. R. Reddy, and L. J. Marnett, Proc. Natl. Acad. Sci. USA **94**, 8652 (1997).

[247] L. J. Marnett and P. C. Burcham, Chem. Res. Toxicol. **6**, 771 (1993).

[248] L. A. VanderVeen, M. F. Hashim, L. V. Nechev, T. M. Harris, C. M. Harris, and L. J. Marnett, J. Biol. Chem. **276**, 9066 (2001).

[249] A. K. Basu, M. L. Wood, L. J. Niedernhofer, L. A. Ramos, and J. M. Essigmann, Biochemistry **32**, 12793 (1993).

[250] C. de los Santos, M. K. andKevin Yarema, A. Basu, J. Essigmann, and D. J. . Patel, Biochemistry **30**, 1828 (1991).

[251] N. L. Morrow, Environ. Health Perspect. **86**, 7 (1990).

[252] N. Tretyakova, R. Sangaiah, T.-Y. Yen, and J. A. Swenberg, Chem. Res. Toxicol. **10**, 779 (1997).

[253] S. Park, C. Anderson, R. Loeber, M. Seetharaman, R. Jones, and N. Tretyakova, J. Am. Chem. Soc. **127**, 14355 (2005).

[254] S. Antsypovich, D. Quirk-Dorr, C. Pitts, and N. Tretyakova, Chem. Res. Toxicol. **20**, 641 (2007).

[255] X.-Y. Zhang and A. A. Elfarra, Chem. Res. Toxicol. **16**, 1606 (2003).

[256] X.-Y. Zhang and A. A. Elfarra, Chem. Res. Toxicol. **17**, 521 (2004).

[257] L. Recio, A.-M. Steen, L. J. Pluta, K. G. Meyer, and C. J. Saranko, Chem. Biol. Interact. **135**, 325 (2001).

[258] W. L. Jorgensen and N. A. McDonald, J. Mol. Struct. (Theochem) **424**, 145 (1998).

[259] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).

[260] R. Aduri, B. T. Psciuk, P. Saro, H. Taniga, H. B. Schlegel, and J. John SantaLucia, J. Chem. Theory Comput. **3**, 1464 (2007).

[261] A. D. MacKerell, Jr., B. Brooks, C. L. Brooks, III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, *Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, III, and P. R. Schreiner, vol. 1, 271–277 (John Wiley & Sons, Chichester, 1998).

[262] N. Foloppe and A. D. MacKerell, Jr., J. Comput. Chem. **21**, 86 (2000).

[263] E. Mayaan, A. Moser, A. D. M. Jr., and D. M. York, J. Comput. Chem. **28**, 495 (2007).

[264] B. C. Stark, R. Kole, E. J. Bowman, and S. Altman, Proc. Natl. Acad. Sci. USA **75**, 3717 (1978).

[265] C. Guerrier-Takada, K. Gardiner, and T. Maresh, Cell **35**, 849 (1983).

[266] T. Cech, A. Zaug, and P. Grabowski, Proc. Natl. Acad. Sci. USA **76**, 5051 (1979).

[267] A. J. Zaug and T. R. Cech, Science **231**, 470 (1986).

[268] W. G. Scott, J. B. Murray, J. R. P. Arnold, B. L. Stoddard, and A. Klug, Science **274**, 2065 (1996).

[269] W. G. Scott, Q. Rev. Biophys. **32**, 241 (1999).

[270] P. B. Rupert, A. P. Massey, S. T. Sigurdsson, and A. R. Ferré-D'Amaré, Science **298**, 1421 (2002).

[271] Adrian R. Ferré-Dámaré, K. Zhou, and J. A. Doudna, Nature **395**, 567 (1998).

[272] I.-h. Shih and M. D. Been, Annu. Rev. Biochem. **71**, 887 (2002).

[273] E. A. Doherty and J. A. Doudna, Annu. Rev. Biophys. Biomol. Struct. **30**, 457 (2001).

[274] Y. Takagi, Y. Ikeda, and K. Taira, Top. Curr. Chem. **232**, 213 (2004).

[275] Y. Takagi and K. Taira, J. Am. Chem. Soc. **124**, 3850 (2002).

[276] Q.-C. He, J.-M. Zhou, D.-M. Zhou, Y. Nakamatsu, T. Baba, and K. Taira, Biomacromol. **3**, 69 (2002).

[277] S. Sawata, M. Komiyama, and K. Taira, J. Am. Chem. Soc. **117**, 2357 (1995).

[278] K. Suzumura, Y. Takagi, M. Orita, and K. Taira, J. Am. Chem. Soc. **126**, 15504 (2004).

[279] V. J. D. Laura M. Hunsicker, J. Inorg. Biochem. **80**, 271 (2000).

[280] E. C. Scott and O. C. Uhlenbeck, Nucleic Acids Res. **27**, 479 (1999).

[281] H. W. Pley, D. S. Lindes, C. DeLuca-Flaherty, and D. B. McKay, J. Biol. Chem. **268**, 19656 (1993).

[282] J. B. Murray, H. Szöke, A. Szöke, and W. G. Scott, Mol. Cell **5**, 279 (2000).

[283] D. L. Beveridge and K. J. McConnell, Curr. Opin. Struct. Biol. **10**, 182 (2000).

[284] P. Auffinger and E. Westhof, Curr. Opin. Struct. Biol. **8**, 227 (1998).

[285] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[286] E. Giudice and R. Lavery, Acc. Chem. Res. **35**, 350 (2002).

[287] T. E. Cheatham, III and M. A.Young, Biopolymers **56**, 232 (2001).

[288] A. D. MacKerell, Jr., J. Comput. Chem. **25**, 1584 (2004).

[289] R. A. Torres and T. C. Bruice, J. Am. Chem. Soc. **122**, 781 (2000).

[290] J. Sarzynska, L. Nilsson, and T. Kulinski, Biophys. J. **85**, 1522 (2003).

[291] Radovan Dvorsky, Josef Sevcik, Leo S. D. Caves, Roderick E. Hubbard, and Chandra S. Verma, J. Phys. Chem. B **104**, 10387 (2000).

[292] K. Réblová, N. Špačková, R. Štefl, K. Csaszar, J. Koča, N. B. Leontis, and J. Šponer, Biophys. J. **84**, 3564 (2003).

[293] M. Boero, K. Terakura, and M. Tateno, J. Am. Chem. Soc. **124**, 8949 (2002).

[294] S.-Y. Le, J.-H. Chen, and N. P. J. V.Maizel, Jr, J. Biomol. Struct. Dyn. **6**, 1 (1998).

[295] P. Auffinger and E. Westhof, J. Mol. Biol. **269**, 326 (1997).

[296] T. Hermann, P. Auffinger, W. G. Scott, and E. Westhof, Nucleic Acids Res. **25**, 3421 (1997).

[297] T. Hermann, P. Auffinger, and E. Westhof, Eur. Biophys. J. **27**, 153 (1998).

[298] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).

[299] A. D. MacKerell Jr, N. Banavali, and N. Foloppe, Biopolymers **56**, 257 (2001).

[300] A. Stone, *The Theory of Intermolecular Forces*, vol. 32 of *International Series of Monographs in Chemistry* (Clarendon Press, Oxford, 1996).

[301] A. D. MacKerell, Jr. and N. K. Banavali, J. Comput. Chem. **21**, 105 (2000).

[302] A. D. MacKerell, Jr., J. Wiórkiewicz-Kuczera, and M. Karplus, J. Am. Chem. Soc. **117**, 11946 (1995).

[303] L. Yang and B. M. Pettitt, J. Phys. Chem. A **100**, 100 (1996).

[304] M. Feig and B. M. Pettitt, J. Phys. Chem. B **101**, 7361 (1997).

[305] http://ndbserver.rutgers.edu/, Nucliec acids database.

[306] Y. Liu, X. Lopez, and D. M. York, Chem. Commun. **31**, 3909 (2005).

[307] L. E. Chirlian and M. M. Francl, J. Comput. Chem. **8**, 894 (1987).

[308] C. M. Breneman and K. B. Wiberg, J. Comput. Chem. **11**, 361 (1990).

[309] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).

[310] V. Atereshko, S. T. Wallace, N. Usman, F. E. Wincott, and M. Egli, RNA **7**, 405 (2001).

[311] G. A. Soukup and R. R. Breaker, RNA **5**, 1308 (1999).

[312] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

[313] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and W. P. Flannery, *Numerical Recipes in Fortran*, 2nd ed. (Cambridge University Press, Cambridge, 1992).

[314] K. J. Hertel and O. C. Uhlenbeck, Biochemistry **34**, 1744 (1995).

[315] M. Koizumi and E. Ohtsuka, Biochemistry **30**, 5145 (1991).

[316] L. A. Cunningham, J. Li, and Y. Lu, J. Am. Chem. Soc. **120**, 4518 (1998).

[317] H. Lee, T. A. Darden, and L. G. Pedersen, J. Chem. Phys. **102**, 3830 (1995).

[318] M. Cossi, G. Scalmani, N. Rega, and V. Barone, J. Chem. Phys. **117**, 43 (2002).

[319] J. B. Murray, D. P. Terwey, L. Maloney, A. Karpeisky, N. Usman, L. Beigelman, and W. G. Scott, Cell **92**, 665 (1998).

[320] M. Oivanen, S. Kuusela, and H. Lönnberg, Chem. Rev. **98**, 961 (1998).

[321] R. G. Pearson, J. Chem. Educ. **64**, 562 (1987).

[322] D.-M. Zhou, Q.-C. He, J.-M. Zhou, and K. Taira, FEBS Lett. **431**, 154 (1998).

[323] K. Yoshinari and K. Taira, Nucleic Acids Res. **28**, 1730 (2000).

[324] J. Florián, M. Štrajbl, and A. Warshel, J. Am. Chem. Soc. **120**, 7959 (1998).

[325] J. B. Murray, C. M. Dunham, and W. G. Scott, J. Mol. Biol. **315**, 121 (2002).

[326] W. G. Scott, Curr. Opin. Struct. Biol. **8**, 720 (1998).

[327] M. Born, Z. Phys. **1**, 45 (1920).

[328] R. Wolfenden and R. Williams, J. Am. Chem. Soc. **105**, 1028 (1983).

[329] H.-J. Werner, P. J. Knowles, R. Lindh, M. Schütz, P. Celani, T. Korona, F. R. Manby, G. Rauhut, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, A. W. Lloyd, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, P. Palmieri, R. Pitzer, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni, and T. Thorsteinsson, MOLPRO, version 2002.6, a package of ab initio programs (2003). See http://www.molpro.net.

[330] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian 98, Revision A.9, Gaussian, Inc., Pittsburgh PA (1998).

[331] J. D. Xidos, J. Li, J. D. Thompson, G. D. Hawkins, P. D. Winget, T. Zhu, D. Rinaldi, D. A. Liotard, C. J. Cramer, D. G. Truhlar, and M. J. Frisch, *MN-GSM: A Module Incorporating the SM5.42 Solvation Models, the CM2 Charge Model, and Löwdin Population Analysis in the Gaussian98 Program*, 1.8 ed. (2002).

[332] E. Cancès, B. Mennucci, and J. Tomasi, J. Chem. Phys. **107**, 3032 (1997).

[333] J. Foresman, T. Keith, K. Wiberg, J. Snoonian, and M. Frisch, J. Phys. Chem. **100**, 16098 (1996).

[334] J. D. Thompson, C. J. Cramer, and D. G. Truhlar, J. Phys. Chem. A. **108**, 6532 (2004).

[335] S. Cabani, P. Gianni, V. Mollica, and L. Lepori, J. Solution Chem. **10**, 563 (1981).

[336] R. C. Pearson, J. Am. Chem. Soc. **108**, 6109 (1986).

[337] J. Florián and A. Warshel, J. Phys. Chem. B **101**, 5583 (1997).

[338] J. Li, T. Zhu, G. D. Hawkins, P. Winget, D. A. Liotard, C. J. Cramer, and D. G. Truhlar, Theor. Chem. Acc. **103**, 9 (1999).

[339] C. Cramer and D. Truhlar, *Rational drug design*, chap. 4, 63–95 (Springer, New York, 1999).

[340] J. R. P. Jr and J. M. Riveros, Phys. Chem. Chem. Phys. **4**, 1622 (2002).

[341] R. Stewart, *The Proton: Applications to Organic Chemistry*, vol. 46 of *Organic Chemistry*, chap. 2, 9–86 (Academic Press, New York, 1985).

[342] D. R. Lide, editor, *CRC handbook of chemistry and physics*, 83 ed. (CRC Press LLC, Boca Raton,FL, 2003).

[343] A. Albert and E. Serjeant, *Ionization Constants of Acids and Bases* (John Wiley and Sons Inc, New York, 1962).

[344] D. D. Perrin, *Ionization constants of inorganic acids and bases in aqueous solution*, 2 ed., IUPAC chemical data series; no 29 (Pergamon Press, New York, 1982).

[345] R. A. Morris, W. B. Knighton, A. A. Viggiano, B. C. Hoffman, and H. F. S. III, J. Chem. Phys. **106**, 3545 (1997).

[346] P. George, R. J. Witonsky, M. Trachtman, C. Wu, W. Dorwart, L. Richman, W. Richman, F. Shurayh, and B. Lentz, Biochim. Biophys. Acta **223**, 1 (1970).

[347] E. M. Evleth, Y. Akacem, and M. Colvin, Chem. Phys. Lett. **227**, 412 (1994).

[348] R. Wolfenden and C. A. Lewis, J. Theor. Biol. **59**, 231 (1976).