The development of Hybrid Quantum Classical Computational Methods for Carbohydrate and Hypervalent Phosphoric systems

KRISHNA KUBEN GOVENDER

Supervisor: Professor Kevin J. Naidoo

Co-supervisor: Doctor Gerhard A. Venter

A thesis submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In the

Scientific Computing Research Unit

Department of Chemistry

University of Cape Town

August 2014

Abstract

Ab initio, density functional theory, and semi-empirical methods serve as major computational tools for quantum mechanical calculations of medium to large molecular systems. Semi-empirical methods are most effectively used in a hybrid quantum mechanics/molecular mechanics (QM/MM) dynamics framework. However, semi-empirical methods have been designed to provide accurate results for organic molecules, but often fail to treat hypervalent species accurately due to their use of an *sp* basis. Recently, significant breakthroughs have been made with the incorporation of *d*-orbitals into the semi-empirical framework, thereby allowing for accurate modeling of both hypervalent and transition metal systems. Here I consider two methods that adopt this new methodology, namely AM1/d-PhoT and AM1*.

Our major focus is the simulation of chemical biological and more specifically chemical glycobiological problems of biochemical interest. When I tested the ability of both AM1/d-PhoT and AM1* to reproduce key metrics in chemical glycobiology (i.e., sugar ring pucker, phosphate participation in transferase reactions) these methods, in combination with the published parameters, performed very poorly. Using the AM1/d-PhoT and AM1* Hamiltonians I set out to re-parameterize these methods aiming to produce holistic biochemical QM/MM toolsets able to simulate fundamental problems of binding and enzyme reactivity in chemical glycobiology. We called these methods AM1/d-CB1 and AM1*-CB1. In the development of these parameter sets I focused specifically on proton transfer, carbohydrate ring puckering, bond polarization, amino acid interactions, and phosphate interactions (facets important to chemical glycobiology). Both AM1/d-CB1 and AM1*-CB1 make use of a *variable property optimization* parameter approach for the glycan molecular class and its chemical environment.

The accuracy of these methods is evaluated for carbohydrates, amino acids and phosphates present in catalytic domains of glycoenzymes, and the are shown to be more accurate for key performance indices (puckering, etc.) and on average across all simulation derived properties (QM/MM polarization, protein performance, etc.) than all other NDDO semi-empirical methods currently being used. A major objective of the newly developed AM1/d-CB1 and AM1*-CB1 is to provide a platform to accurately model reactions central to chemical glycobiology using hybrid QM/MM molecular dynamics (MD) simulations. AM1/d-CB1 is applied to a well-known reaction involving purine nucleoside phosphorylase (PNP) and results

lead me to conclude that the method shows promise for modelling glycobiological QM/MM systems.

Declaration

I declare that the work in this thesis, "The development of Hybrid Quantum Classical Computational Methods for Carbohydrate and Hypervalent Phosphoric systems" is based on research conducted at the Scientific Computing Research Unit, Department of Chemistry, University of Cape Town. No part of this thesis has been submitted elsewhere for any other degree or qualification. All work in the text is my own, unless otherwise stated.

Krishna Kuben Govender

August 2014

Acknowledgements

I am thankful for financial assistance provided by the University of Cape Town, Department of Chemistry, Equity Development Program (EDP); South African Research Chair Initiative (SARChI); National Research Foundation (NRF) Innovation award.

Thank you to my supervisors, Professor Kevin J. Naidoo and Doctor Gerhard J. Venter, as well as Professor Jiali Gao (University of Minnesota) for allowing me the opportunity to work with you and helping me better understand the various aspects of this complex research field. Your knowledge, encouragement and vast discussions are what helped me complete this project successfully.

I would like to extend a wholehearted appreciation to my loving wife (Jestine) for always being so supportive and providing me with that much needed boost of enthusiasm. Other family members who deserve tremendous praise are my loving parents and siblings who have always been there for me financially and emotionally. My two little girls (Xara and Dharshnee) deserve a special thank you for always keeping me on my toes.

Thank you to members of the Scientific Computing Research Unit for assisting with various problems experienced during the course of this project. I would like to extend further thanks to Dr. Chris Barnett, Dr. Riedaa Gamieldian, Kyle Fernandes and Ian Rogers for providing me with much needed scripts that helped make the analysis of data quick and painless.

Finally, thank you to Louise Bezuidenhout, the administrative staff member of the Scientific Computing Research Unit; for always making sure that my administrative documents are processed in an efficient manner and for providing some opportune advice.

List of Abbreviations

Abbreviation	Definition
DNA	Deoxyribonucleic acid
GT	Glycosyltransferase
NMR	Nuclear Magnetic Resonance
AFM	Atomic Force Microscope
MM	Molecular Mechanics
LJ	Lennard-Jones
DFT	Density Functional Theory
SE	Semi-empirical
HMO	Hückel Molecular Orbital
PPP	Pariser-Parr-Pople
EHT	Extended Hückel Theory
ZDO	Zero Differential Overlap
CNDO	Complete Neglect of Differential Overlap
INDO	Intermediate Neglect of Differential Overlap
ZINDO	Zerner Intermediate Neglect of Differential Overlap
SINDO	Symmetrically Orthogonalized Intermediate Neglect of Differential Overlap
NDDO	Neglect of Diatomic Differential Overlap
MINDO	Modified Intermediate Neglect of Differential Overlap
ΔH_f	Heat of formation

MNDO	Modified Neglect of Differential Overlap
AM1	Austin Model 1
PM3	Parametric Method Number 3
MNDO/d	Modified Neglect of Differential Overlap with *d*-orbitals included
SAM1	Semi *ab initio* method 1
PM3(tm)	Parametric Method Number 3 with *d*-orbitals for transition metals
AM1/d	Austin Model 1 with *d*-orbitals included
GCF	Gaussian Core Function
PIF	Parameterized Interaction Function
MAIS	Method Adapted for Intermolecular Studies
PDDG	Pairwise Distance Directed Gaussian
AM1*	Austin Model 1 with the inclusion of *d*-orbitals
PM3CARB-1	Parametric Method Number 3 with carbohydrate specific parameters
$PM3^{MS}$	Parameteric Method Number 3 for monosaccharides
QM	Quantum Mechanics
SRP	Specific Reaction Parameter
RM1	Recife Model 1
AM1/d-PhoT	Austin Model 1 with *d*-orbitals and specific parameters for phosphoryl transfer reactions
QM/MM	Quantum mechanics/Molecular mechanics
PM6	Parametric Method Number 6
AUE	Average unsigned error

OM1	Orthogonalization Model 1
OM2	Orthogonalization Model 2
OM3	Orthogonalization Model 3
AM1-D	Austin Model 1 with a correction for dispersion
PM3-D	Parametric Method Number 3 with a correction for dispersion
DH	First generation dispersion and hydrogen bond corrections
DH2	Second generation dispersion and hydrogen bond corrections
DH+	Third generation dispersion and hydrogen bond corrections
D3H4	Fourth generation dispersion and hydrogen bond corrections
PM7	Parametric Method Number 7
SCC-DFTB	Self-consistent charge density functional tight-binding
LCAO	Linear Combination of Atomic Orbital
vdW	van der Waals
LSCF	Local Self Consistent Field
SLBO	Strictly localized bond orbital
GHO	Generalized Hybrid Orbital
MD	Molecular Dynamics
HF	Hartree–Fock
SCF	Self-consistent-field
LCAO	Linear combination of atomic orbitals
LDA	Local density approximation
LYP	Lee, Yang, Parr

PW	Perdew-Wang
VWN	Vosko, Wilk, Nusair
B3LYP	Becke's three parameter functional including LYP correlation
LSDA	Local spin density approximated
GGA	Generalized gradient approximation
VSXC	van Voorhis and Scuseria's τ-dependent gradient-corrected correlation
ps	picoseconds
GA	Genetic algorithm
MSE	Mean signed error
MUE	Mean unsigned error
VPO	Variable property optimization
AM1/d-CB1	Austin Model 1 with the inclusion of *d*-orbitals and parameters for chemical biological systems
AM1*-CB1	Austin Model 1 with *d*-orbitals and parameters for chemical biological systems
FEARCF	Free Energy from Adaptive Reaction Coordinate Forces
HB	Hydrogen bond
PNP	Purine Nucleoside Phosphorylase
dGTP	deoxy-guanosine-triphosphate
TS	transition state

Table of Contents

List of Figures

List of Tables

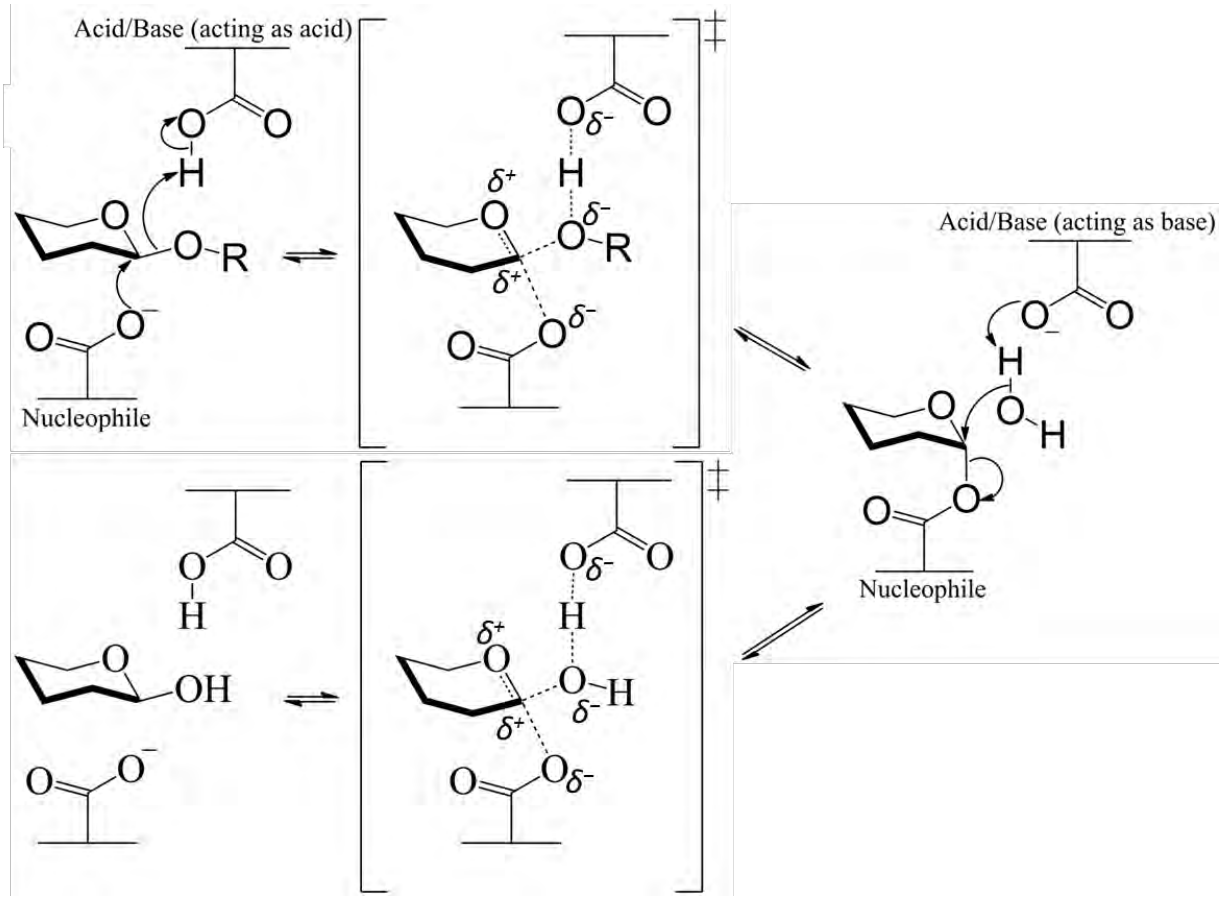
1. Introduction

A brief introduction into the field of chemical glycobiology is provided. This is followed by a detailed description of monosaccharides and the structural complexity that they possess. Conformational analysis of monosaccharides is presented with respect to theoretical and experimental developments in the field. The chapter is concluded with the objectives of the current work.

The field of glycobiology highlights the myriad of complex processes in which carbohydrates play a vital role.[1] Glycans, in the form of oligosaccharides, polysaccharides, glycoproteins, glycolipids, proteoglycans, and other glycoconjugates, can be key players in a number of important biological recognition processes, such as: intercellular trafficking, cell adhesion development, cancer progression, host-pathogen interaction, and immune response, just to name a few.[1-5] Nucleic acids can be made easily and cheaply via chemical and biological synthetic techniques, and protein sequences, which are encoded by DNA, can be easily determined, produced, and manipulated through recombinant DNA technology. Unlike proteins and nucleic acids, glycans are generally more difficult to synthesize because the molecules are typically branched rather than linear, and the monosaccharide units making up the glycan can be connected via α or β linkages.[6] This makes glycans considerably complex, involving various types of sugar processing and trimming under the action of a series of competing enzymes along secretory pathways.[1] However, this structural complexity offers powerful opportunities in the design of molecular experiments that will unpick the mechanism of this biology. In this way chemistry offers one unique, and as yet largely unrealized, strategy for dissecting this complexity where strictly biological approaches may fail.[1] The increased appreciation for the ubiquity of glycans and their importance to human health has spawned the field of chemical glycobiology.[7]

Glycosyltransferases (GTs) constitute a large family of enzymes that are involved in the biosynthesis of glycans.[8] GTs are highly regio- and stereo- selective enzymes and have been successfully applied for enzymatic synthesis of oligosaccharides, which can be isolated from milk, serum, and organ tissues.[9] Particularly abundant are the GTs that transfer a sugar residue from an activated nucleotide sugar donor, to specific acceptor molecules, forming glycosidic bonds.[10] Transfer of the sugar residue occurs via one of two categories: *retaining*, where the

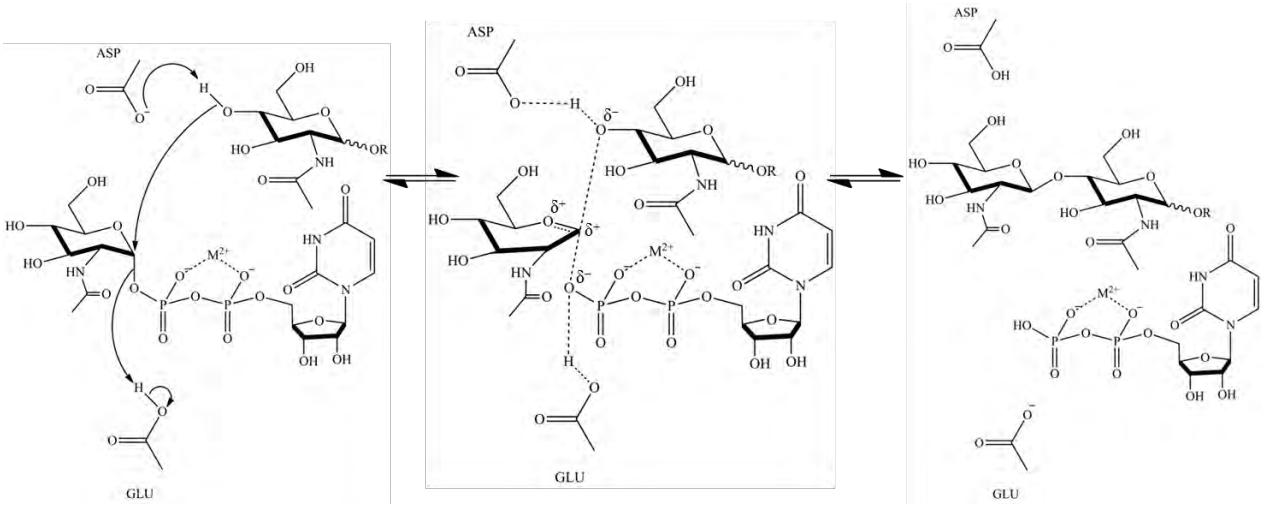
product glycoside has the same stereochemistry as the activated leaving group, and *inverting*, where the reaction proceeds with inversion of the stereochemistry at the anomeric carbon. These enzymes are present in both prokaryotes and eukaryotes, and generally display exquisite specificity for both the glycosyl donor and the acceptor substrates.[10] An example of a reaction which proceeds via a retaining mechanism is provided in Scheme 1.1 in which a double displacement occurs resulting in retention of the stereochemistry at the anomeric carbon.



Scheme 1.1: General mechanism for a retaining glycosidase.

Scheme 1.2 illustrates the mechanistic pathway followed by an inverting glycosyltransferase-catalyzed reaction. What is interesting is that both the retaining and inverting mechanisms produce an oxocarbenium ion-like transition state (TS) where the ring puckers into a conformation other than the favorable $^{1}C_{4}$ or $^{4}C_{1}$ chair, which is an important role player in glycosidase and glycotransferase reactions. Stabilization of a positively charged oxocarbenium

ion is therefore a potential catalytic strategy, and some enzymes appear to act primarily by directly stabilizing the oxocarbenium ion.[11-14] Clearly the carbohydrate moiety and the puckering which it undergoes are important facets within the chemical glycobiological framework and in the sections that follow we shall explore these systems in more detail.



Scheme 1.2: Mechanism for inverting glycosyltransferase reaction involving a UDP-GlcNAc donor and GlcNAc-containing acceptor substrates.

1.1 Carbohydrates

Carbohydrates are important molecules in biological systems.[15] They are energy stores, fuels, and metabolic intermediates; they are found in the back-bone of nucleic acids and in cell walls of bacteria and plants (*cellulose* is one of the most abundant polysaccharide compounds in the biosphere); they interact and link with other macromolecules to form a wide range of glycoconjugates which are determinant for cell-cell communication and interaction between cells and other elements in their environment;[16] they are key role players in chemical glycobiology, such as in glycosyltranferase reactions for example (as shown above). In addition, carbohydrates have several industrial applications such as: starch in manufacturing of goods and pastas, gum in food processing, mono and oligo-saccharides as sweetners; cotton and linen in clothing fabrics.[16]

The most basic carbohydrate unit is the *monosaccharide*, which has many degrees of freedom, implying that there are various conformations accessible to this moiety. Therefore,

even the most basic carbohydrate (*monosaccharide*) is chemically complex, possessing diverse functionality.

1.1.1 Monosaccharides

Monosaccharides are the building blocks of carbohydrates. Hence, the general characteristics of the latter naturally depend on the structural properties of the former. Monosaccharides generally fall into two categories; hexoses and pentoses. Hexoses contain six carbon atoms while pentoses contain five carbon atoms. These molecules are not linear and prefer to cyclize and form an intramolecular acetal.[17] The six membered ring formed from the cyclization is termed a pyranose and the five membered ring is a furanose. An example of the cyclization of glucopyranose and ribofuranose is provided in Figure 1.1 below.

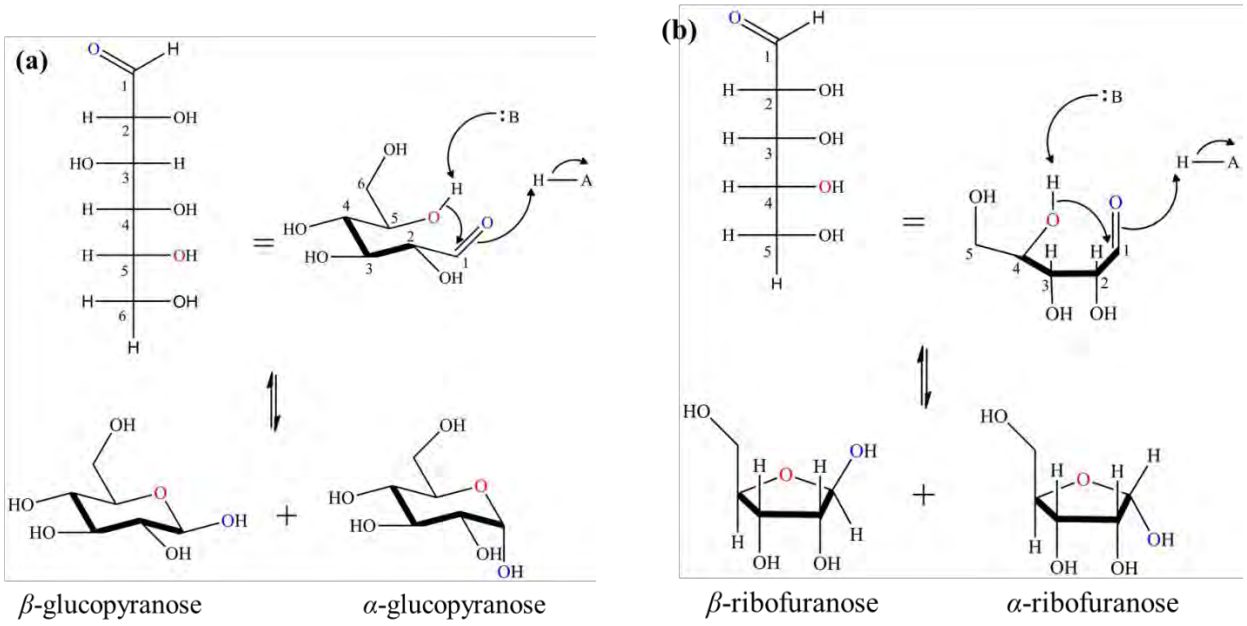


Figure 1.1: The monosaccharide units of sugars in the linear (open) form with aldehyde and alcohol functionality. Intramolecular acetal formation occurs to yield a closed cyclic form for (a) glucopyranose and (b) ribofuranose.

Upon cyclization a new chiral center is formed at C1, typically referred to as the anomeric carbon. When the hydroxyl group points downwards (below the plane of the ring) it is in the α position and when it points upwards it is in the β position. These forms are called α- and

β-anomers and either is possible after a cyclization. Depending on the electronegative substituents at the anomeric position of, for example pyranoses, there is a tendency for the axial configuration.[17] This is known as the anomeric effect[18] and provides stabilization of the axial substituent such that the inherent steric bias of the substituent is overcome.[17] It has been stated, in the electrostatic model,[19] that there is an increased preference for the electronegative group to be axial due to the repulsive dipole-dipole interactions.[20] Figure 1.2 shows that an equatorial C-X bond has the C-X and net C-O dipole moments pointing in the same direction. This causes the dipoles to be additive, destabilizing the molecule and thus increasing the energy. When the C-X bond is axial the dipole moments partially cancel out, minimizing the destabilization, thereby causing a more stable conformation.

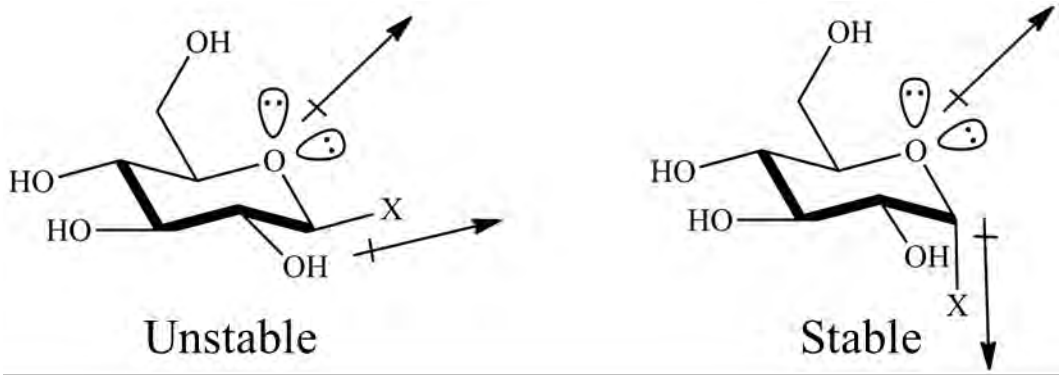


Figure 1.2: Dipole moments of a halogenated derivative of glucopyranose (X = Halogen). Equatorial conformer is provided on the left and axial on the right.

1.1.2 Linking monosaccharides

Monosaccharide units have several alcohol groups which can undergo a condensation reaction to form a glycosidic linkage between various monosaccharides.[17] Two monosaccharide units which are glycosidically linked are referred to as disaccharides. Three to ten glycosidically linked monosaccharides are termed oligosaccharides, while more than ten linked units are referred to as polysaccharides. Due to the various number of hydroxyl species located along the periphery of the monosaccharide (Figure 1.1), numerous different types of linkages are possible. Figure 1.3 illustrates this variety by providing examples of a few commonly occurring, glycosidically linked, molecules. Sucrose (table sugar) is formed via an ether bond between C1

of glucose and C2 of fructose. The bond is formed in such a way that the glycosidic linkage lies α to the C1 carbon. This is termed a α (1→2) glycosidic bond. Lactulose and lactose (milk sugar) have glycosidic linkages between galactose and fructose, and galactose and glucose, respectively. In both cases the linkage lies β to the C1 carbon thereby being termed a β (1→4) glycosidic linkage. Maltose and cellobiose are α and β glycosidically linked disaccharides comprised of two glucose monomers. Cellobiose can be further linked to form cellulose, a structural polymer found in plants. The nucleoside adenosine comprises a base and a furanose sugar (ribose). A nitrogen atom (N9) of the base is N-glycosidically linked to the C1 atom of ribose.

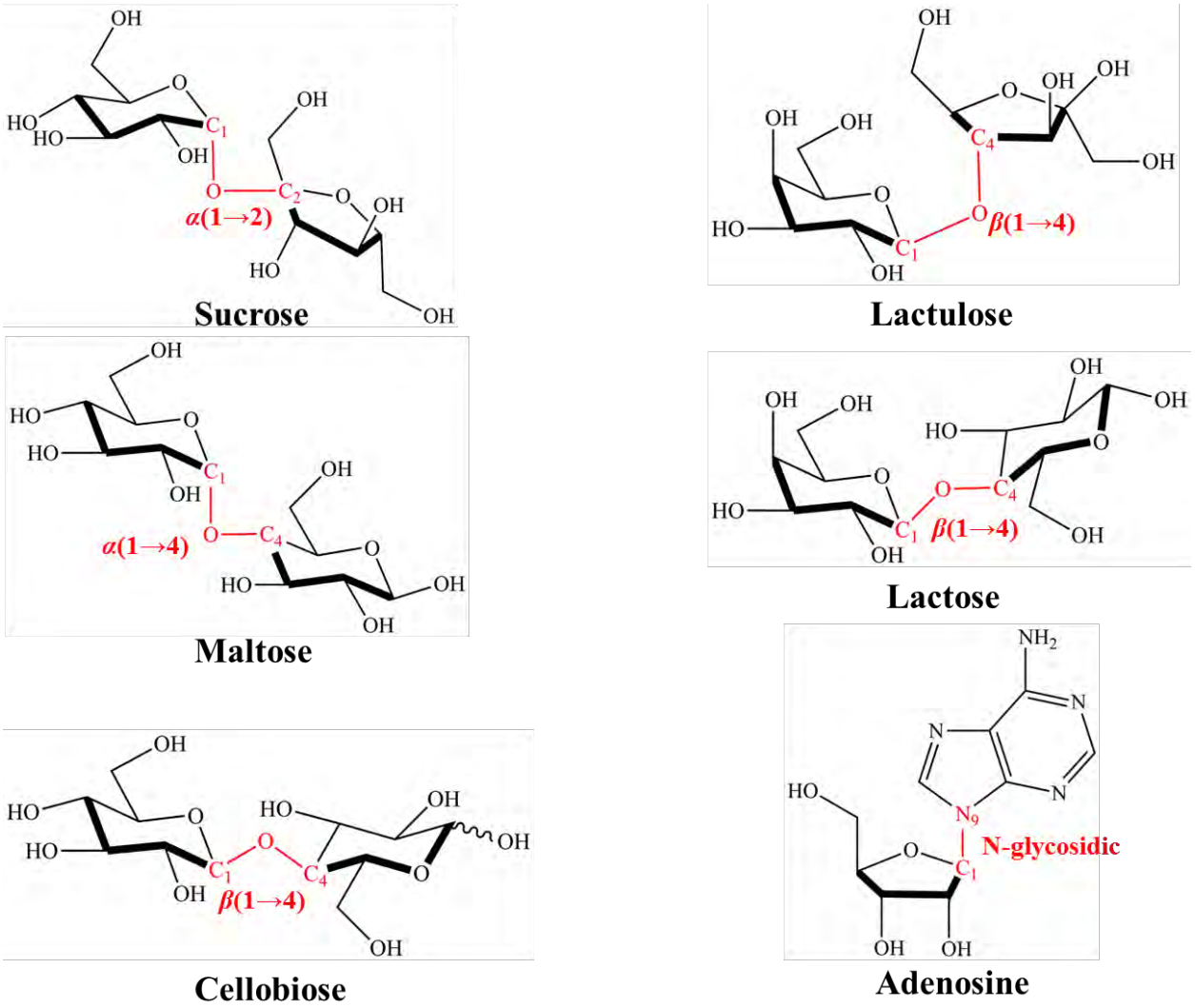


Figure 1.3: Some common glycosidic linkages (highlighted in red).

1.1.3 Pucker

The monomers of sugars, being cyclic molecules, are not stable as planar rings, but instead adopt a range of ring conformations or *puckers*. The puckering of cyclic molecules has been subject to investigation for more than a century,[21,22] with the existence of various ring puckering conformers being put forward by Haworth.[23] This motion occurs due to two main reasons:

i) Atoms linked to carbons prefer a tetrahedral spatial distribution if possible.
ii) Ring substituents hinder each other if they are in close contact (e.g. hydroxyl groups on the same side of the ring).

The reasons provided above imply that steric and stereo-electronic hindrance allows/imposes peculiar spatial arrangements to minimize or, if possible, eliminate strain effects.[16] These strain effects can be broken down into two categories:

i) *Rotamers* which are ring substituent orientations corresponding to minima, with respect to strain hindrance.
ii) *Conformers* which are ring forms corresponding to minima, with respect to chain strains.

An interconversion between rotamers and conformers occur only by means of rotation around torsion angles.

The primary alcohol group (-CH_2OH) of a monosaccharide has a preference to form a torsional angle with the ring oxygen (O-C-C-O) such that the oxygens' adopt a synclinal (gauche) conformation. The tendency for the torsional angle about C-X-Y-C or X-C-C-Y molecular fragments (where X and Y are electronegative atoms) to prefer the gauche conformations is characteristic of a rotamer, termed the gauche effect.[16,17]

For conformers it is possible to calculate, theoretically, a set of *ideal conformations* for a generic *N*-membered ring structure. In ideal structures, the special orientation for (almost) all substituents at every carbon atom follows an exact tetrahedral arrangement. For six-membered rings there are 38 ideal conformers,[24-26] divided into stable conformers (*chairs*, *boats*, and *skew-boats*) and transition state conformers (*half-chairs* and *envelopes*). The various conformers are illustrated in Figure 1.4. There is less variability for five-membered rings because there are only

two possible groups of ideal structures, divided into stable (*envelope*) and metastable (*twist*) forms. This produces only 20 ideal conformers for the five-membered rings, which are illustrated in Figure 1.5. The interconversion between ring forms, in this case, is simpler because both twist and envelope conformers are flexible.[16]

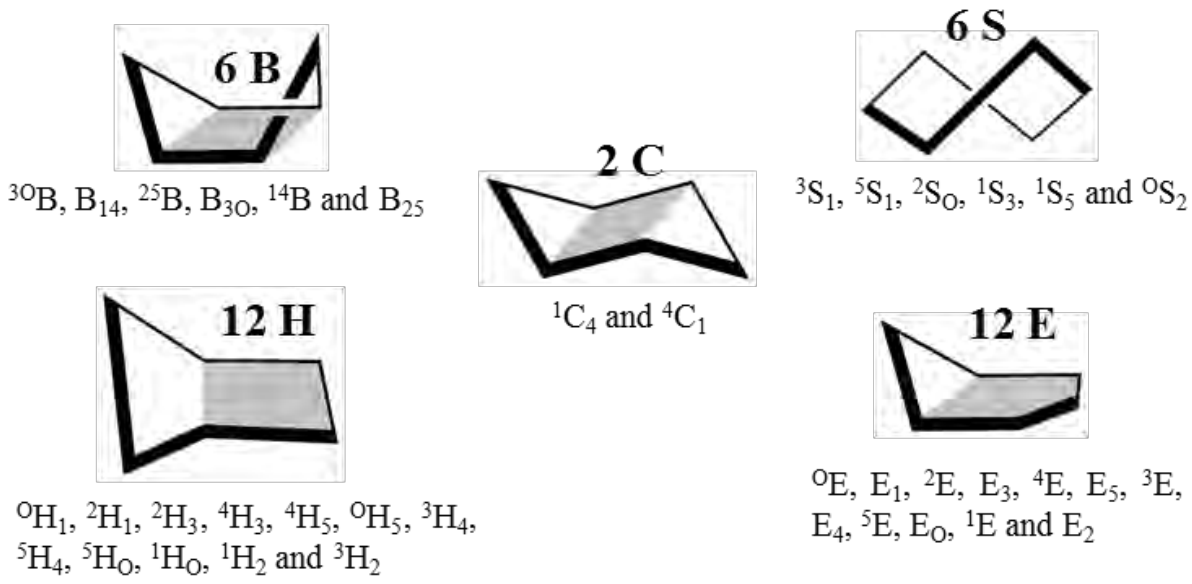


Figure 1.4: The 38 canonical conformers available to a six-memebered ring: Two chairs (C), six boats (B), six skew-boats (S), twelve half-chairs (H), and twelve envelopes (E). The shaded area connects atoms in the same plane. Names are listed according to IUPAC recommendation.[25]

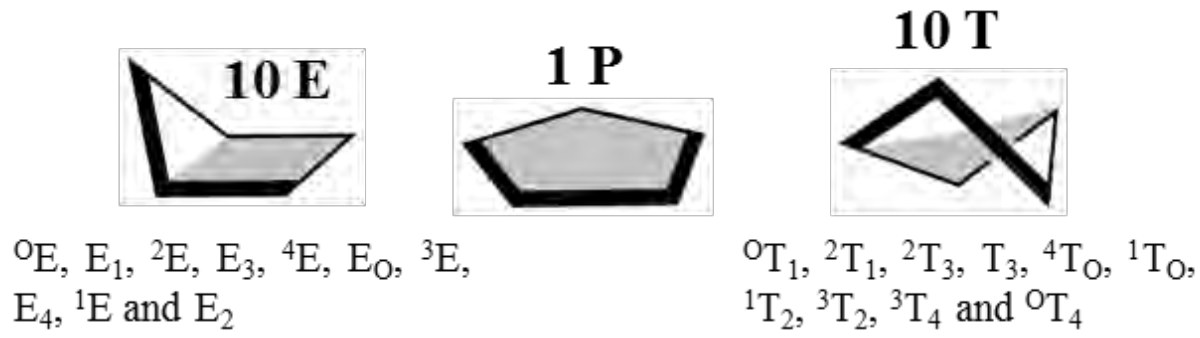


Figure 1.5: The 21 canonical conformers available to a five-membered ring. 10 envelopes (E), 10 twists (T), and 1 planar (P). The shaded area connects atoms in the same plane. Conformational names are listed according to IUPAC recommendation.[25]

1.1.4 Experimental determination of pucker

The practical methods for examining monosaccharide conformations are X-ray crystallography, solution NMR, and solid state NMR.[27] High-resolution X-ray crystallography gives information on the molecular structure. Solution state NMR, in contrast, gives indirect information on the torsional angles through chemical shifts,[28-31] scalar J-couplings,[32,33] and cross-correlated relaxation effects.[34,35] These techniques encounter difficulties when dealing with very large systems, such as polynucleotide-protein complexes, due to imperfect crystallization and slow molecular rotation resulting in spectral line broadening. Solid state NMR is a technique which does not require long-range crystallinity or rapid molecular motion and thereby circumvents the problems experienced by the other techniques.[27]

Studies of pyranoses such as α-glucose,[36,37] β-glucose,[38,39] cellobiose,[40] and maltose,[41] conducted with the aid of the techniques given above (X-ray crystallography and solid-state NMR) show the $^4C_1$ conformer in the solid state. For furanoses, in the solid state, there is little tendency for the planar conformer, as free sugars and nucleotides prefer envelopes.[42-44] The furanoid sugars are considered to have C_2 or C_3 "meta" to the ring oxygen,[45] i.e. to be in the E_2, ^2E, E_3, or ^3E envelope conformers. A combination of molecular mechanics and NMR has been used to further understand the puckering conformations.[46,47]

Both NMR and X-ray crystallography are only able to distinguish between the chair and inverted chair structures of pyranoses. This is merely because chairs are rigid structures while boats and skew-boats are flexible, rapidly exchanging in solution, making them very difficult to crystallize. In such a case, experimental techniques like *Atomic Force Microscope* (AFM) spectroscopy are able, to some extent, detect flexible conformers. Marszalek et al. used AFM spectroscopy to establish the free energy of glucopyranose boats in the polysaccharide amylose.[48,49] Unlike NMR, researchers make use of polymers as the elongation process can produce various conformers along the polymer in direct relation with the linkage scheme of the polymer itself.[16] Thus, by subtracting free energy differences of different processes the free energy contribution of the sole conformational transition can be evaluated (e.g., skew-boat free energy was estimated to about 25 kJ/mol over the chair conformer).[49]

1.1.5 Theoretical determination of pucker

The mathematical definition of various five- and six-membered ring conformers was based on a spherical polar coordinate system which was introduced by Cremer and Pople in their 1975 publication.[50] These definitions proved very popular despite the cumbersome relationship between the coordinates and physically meaningful stresses and strains on the rings. A more recent definition, and one that is utilized in this thesis, is the triangular decomposition coordinate set proposed by Hill and Reilly.[51] This method enables a ready description of ring conformers as a function of triangular planes deviating from a reference plane placed on the five- or six-membered monosaccharide ring.[51] For an *N*-membered ring there are *N*-2 planes, a central reference plane, and *N*-3 rotatable puckering planes with respect to the central plane.[51,52] For example a five membered ring has a central plane and two puckering planes such that there are two puckering coordinates, θ_0, $\theta_1 \in [-90, 90]$. The angle of pucker is calculated using,

$$\theta_i = \frac{\pi}{2} - \cos^{-1}\left[(q_i \bullet n) \bullet \left(\|q_i\| \bullet \|n\|\right)\right], \tag{1.1}$$

where q_i is the vector normal to the rotatable plane *i* and the axis about which the plane rotates, while *n* is the vector normal to the reference plane. The coordinates are provided in Figure 1.6 for both the five- and six-membered monosaccharide.

Figure 1.6: Reference plane and rotatable planes for (a) five-membered monosaccharide and (b) six-membered monosaccharide, as defined by the method of triangular decomposition/tessellation.

The reference plane used in the triangular decomposition/tessellation[51] is *not* the same as the mean plane used in the Cremer-Pople coordinates.[50]

1.2 Objectives

To date the investigation of reactions which take place in carbohydrate processing enzymes, via experiment, has limited possibilities. These range from X-ray structural analysis of mutated enzymes or enzymes bound to inhibitors,[53,54] to kinetic isotope effect (KIE) experiments.[55] The nature of the TS remains inaccessible to experimentalists. As a result theoreticians make use of computational methods, specifically hybrid quantum classical (QM/MM) methods, to try and better understand the enzyme reaction mechanisms as well as the conformational and electronic nature of the TS.[56] Researchers often treat the quantum mechanical region, of QM/MM simulations, with semi-empirical (SE) methods due to their inherent speed. However, an important aspect that needs to be kept in mind, when using a SE method to model systems important in glycobiology, is that the method must be able to accurately model the conformational and electronic transitions of both furanose and pyranose monosaccharides, or the mechanistic details and TS predicted with the method are meaningless. The aim of this work is to:

i) Re-parameterize currently existing SE methods to accurately model systems that are important in chemical glycobiology.

ii) Test the newly developed methods to ensure that they conserve the stereoelectronic preferences of saccharides as well as model interactions of amino acids and phosphates.

iii) Apply the new methods to hybrid QM/MM MD simulations focusing on enzymatic glycosyl reactions.

1.3 References

(1) Wang, L.-X.; Davis, B. G. *Chem. Sci.* **2013**, *4*, 3381.
(2) Helenius, A.; Aebi, M. *Science* **2001**, *291*, 2364.
(3) Hart, G. W.; Copeland, R. J. *Cell* **2010**, *143*, 672.
(4) Bertozzi, C. R.; Kiessling, L. L. *Science* **2001**, *291*, 2357.
(5) Dwek, R. A. *Chem. Rev.* **1996**, *96*, 683.
(6) Sears, P.; Wong, C.-H. *Science* **2001**, *291*, 2344.
(7) Kiessling, L. L.; Splain, R. A. *Annu. Rev. Biochem.* **2010**, *79*, 619.
(8) Taniguchi, N.; Honke, K.; Fukuda, M. *Handbook of Glycosyltransferase and Related Genes.*; Springer: Tokyo, 2002.
(9) Boons, G.-J.; Demchenko, A. V. *Chem. Rev.* **2000**, *100*, 4539.
(10) Breton, C.; Šnajdrová, L.; Jeanneau, C.; Koča, J.; Imberty, A. *Glycobiology* **2006**, *16*, 29R.

(11) Burkart, M. D.; Vincent, A.; Duffels, B. W.; Ley, S. V.; Wong, C.-H. *Bioorg. Med. Chem.* **2000**, *8*, 1937.
(12) Murray, B. W.; Takayama, S.; Schultz, J.; Wong, C.-H. *Biochemistry-Us* **1997**, *36*, 823.
(13) Hartman, M. C. T.; Jiang, S.; Rush, J. S.; Waechter, C. J.; Coward, J. K. *Biochemistry-Us* **2007**, *46*, 11630.
(14) Berti, P. J.; McCann, J. A. B. *Chem. Rev.* **2006**, *106*, 506.
(15) Kurihara, Y.; Ueda, K. *Carbohy. Res.* **2009**, *344*, 2266.
(16) Autieri, E. PhD Thesis, Universià degli studi di trento, 2011.
(17) Barnett, C. B. PhD Thesis, University of Cape Town, 2010.
(18) Kirby, A. J. *The anomeric effect and related stereoelectronic effects at oxygen*; Springer-Verlag: Heidelberg, 1983.
(19) Edward, J. T. *Chem. Ind.* **1955**, *36*, 1102.
(20) Juaristi, E.; Cuevas, G. *Tetrahedron* **1992**, *48*, 5019.
(21) Juaristi, E. *Conformational Behaviour of Six-Membered Rings: Analysis, Dynamics and Stereoelectronic effects*; VCH Publishers Inc., 1995.
(22) Tipson, R. S. *Advances in Carbohydrate Chemistry and Biochemistry*; Academic Press Inc.: New York, 1971.
(23) Haworth, W. N. *The constitution of sugars*; Arnold: London, 1929.
(24) Schwarz, J. C. P. *J. Chem. Soc., Perkin Trans. 1* **1973**, X002.
(25) Dixon, B. F.; Jeannin, Y.; Loening, K. L.; Moss, G. P. *Eur. J. Biochem.* **1980**, *111*, 295.
(26) McNaught, A. D. *Pure Appl. Chem.* **1996**, *68*, 1919.
(27) van Dam, L.; Ouwerkerk, N.; Brinkmann, A.; Raap, J.; Levitt, M. H. *Biophys. J.* **2002**, *83*, 2835.
(28) Santos, R. A.; Tang, P.; Harbison, G. S. *Biochem.* **1989**, *28*, 9372.
(29) Gorenstein, D. G. *Methods Enzymol.* **1992**, *211*, 254.
(30) Xu, X.-P.; Chiu, W.-L. A. K.; Au-Yeung, S. C. F. *J. Amer. Chem. Soc.* **1998**, *120*, 4230.
(31) Rossi, P.; Harbison, G. S. *J. Magn. Res.* **2001**, *151*, 1.
(32) Davies, D. B. *Prog. Nucl. Mag. Reson. Spectrosc.* **1978**, *12*, 135.
(33) Ippel, J. H.; Wijmenga, S. S.; de Jong, R.; Heus, H. A.; Hilbers, C. W.; de Vroom, E.; van der Marel, G. A.; van Boom, J. H. *Magn. Reson. Chem.* **1996**, *34*, S156.
(34) Boisbouvier, J.; Brutscher, B.; Pardi, A.; Marion, D.; Simorre, J.-P. *J. Amer. Chem. Soc.* **2000**, *122*, 6779.
(35) Felli, I. C.; Richter, C.; Griesinger, C.; Schwalbe, H. *J. Amer. Chem. Soc.* **1999**, *121*, 1956.
(36) McDonald, T. R. R.; Beevers, C. A. *Acta. Cryst.* **1952**, *5*, 654.
(37) Brown, G. M.; Lev, H. A. *Science* **1965**, *147*, 1038.
(38) Ferrier, W. G. *Acta. Cryst.* **1960**, *13*, 678.
(39) Ferrier, W. G. *Acta. Cryst.* **1963**, *16*, 1023.
(40) Chu, S. S. C.; Jeffery, G. A. *Acta. Cryst.* **1968**, *B24*, 830.
(41) Quigley, G. J.; Sarko, A.; Marchessault, R. H. *J. Am. Chem. Soc.* **1970**, *5*, 834.
(42) Sundaralingam, M. *J. Am. Chem. Soc.* **1965**, *87*, 599.
(43) Sundaralingam, M. *Biopolymers* **1969**, *7*, 821.
(44) Spencer, M. *Acta. Cryst.* **1959**, *12*, 59.
(45) Lemieux, R. U.; Nagarajan, R. *Can. J. Chem.* **1964**, *42*, 1270.
(46) Xu, Q.; Bush, C. A. *Biochem.* **1996**, *35*, 14521.

(47) Raap, J.; Van Boom, J. H.; Van Lieshout, H. C.; Haasnoot, C. A. G. *J. Amer. Chem. Soc.* **1988**, *110*, 2736.
(48) Marszalek, P. E.; Oberhauser, A. F.; Pang, Y. P.; Fernandez, J. M. *Nature* **1998**, *396*, 661.
(49) Zhang, Q.; Jaroniec, J.; Lee, G.; Marszalek, P. E. *Angew. Chem. Int. Ed.* **2005**, *44*, 2723.
(50) Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975**, *97*, 1354.
(51) Hill, A. D.; Reilly, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 1031.
(52) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243.
(53) Sulzenbacher, G.; Driguez, H.; Henrissat, B.; Schülein, M.; Davies, G. J. *Biochemistry-Us* **1996**, *35*, 15280.
(54) Vasella, A.; Davies, G. J.; Böhm, M. *Curr. Opin. Chem. Biol.* **2002**, *6*, 619.
(55) Werner, R. M.; Stivers, J. T. *Biochemistry-Us* **2000**, *39*, 14054.
(56) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.

2. Methods

This chapter introduces various computational methods based on classical mechanics (or molecular mechanics) and quantum mechanics. The quantum mechanics sections presented are primarily based on the semi-empirical framework. The theory related to these and other quantum mechanical methods are provided in Chapter 3. In addition some background and theory is provided for hybrid quantum classical QM/MM methods as well as their application and a few aspects to consider when combing them with molecular dynamics.

2.1 Molecular Mechanics

Molecular mechanics (MM)[1,2] is based on a mathematical model of a molecule as a collection of balls (corresponding to the atoms) held together by springs (corresponding to the bonds) (Figure 2.1). The model is conceptually very close to the intuitive feel for molecular energetics that one obtains when manipulating molecular models of plastic or metal: the model resists distortions from the "natural" geometry that corresponds to the bond lengths and angles imposed by the manufacturer, and in the case of space-filling models, atoms cannot be forced too closely together.[3] The principle behind MM is to express the energy of a molecule as a function of its resistance toward bond stretching, bond bending and atom crowding. This energy equation is used to find bond lengths, angles and dihedrals corresponding to the minimum-energy geometry – or more precisely, the various possible potential energy surface minima. The form of the mathematical expression used for the energy, and the parameters in it, constitute a *force field*, and it is for this reason that MM methods are also known as *force field* methods.[3]

Figure 2.1: Line structure of methane (left) and molecular mechanics depiction of methane as a collection of balls (the atoms) held together by springs (the bonds) (right).

In MM the potential energy of the system under consideration is expressed as an analytical function of the 3N coordinates of the N atoms present.[4] The total MM energy is expressed as the sum of bonded and non-bonded energy terms,

$$E_{MM} = E_b + E_{nb}, \tag{2.1}$$

where E_{MM} is the total energy, E_{b} the bonded energy and E_{nb} the non-bonded energy.

E_{b} in the CHARMM force field consists of the following terms,

$$E_b = E_{bond} + E_{angle} + E_{UB} + E_{dihedral} + E_{improper} + E_{CMAP}, \tag{2.2}$$

which is expanded as,

$$\begin{aligned} E_b = & \sum_{bonds} K_b \left(b - b_0\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{Urey-Bradley} K_{UB} \left(S - S_0\right)^2 + \\ & \sum_{dihedrals} K_\varphi \left(1 + \cos\left(n\varphi - \delta\right)\right) + \sum_{impropers} K_\omega \left(\omega - \omega_0\right)^2 + \sum_{residues} u_{CMAP} \left(\Phi, \Psi\right), \end{aligned} \tag{2.3}$$

where K_{b}, K_θ, K_{UB}, K_φ, and K_ω are the bond, angle, Urey–Bradley, dihedral angle and improper dihedral angle force constants, respectively. b, θ, S, φ and ω are the bond length, bond angle, Urey–Bradley distance between atoms 1 and 3, dihedral angle and improper torsion angle, respectively, with the subscript zero representing the equilibrium values for the individual terms. The CMAP term is a cross term for the Φ, Ψ (backbone dihedral angle) values used to treat conformational properties of protein backbones.[5]

E_{nb} is used to model the non-bonded interactions *via* Lennard-Jones (LJ) 6–12 and Coulomb terms,

$$E_{nb} = \sum_{i,j} \varepsilon_{ij} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{6} \right] + \sum_{i,j} \frac{q_i q_j}{\varepsilon_1 r_{ij}}, \tag{2.4}$$

where $\varepsilon_{\mathrm{ij}}$ is the LJ well depth, $R_{ij}^{\min}$ is the distance at the LJ minimum, q_{i} and q_{j} are the partial atomic charges, r_{ij} is the distance between atoms i and j and ε_1 is the effective dielectric constant.

The LJ parameters ($\varepsilon_{\mathrm{ij}}$ and $R_{ij}^{\min}$) between pairs of different atoms are obtained from the Lorentz–Berthelot combination rules,[6] in which $\varepsilon_{\mathrm{ij}}$ values are based on the geometric mean of ε_{i} and ε_{j},

$$\varepsilon_{ij} = \sqrt{\varepsilon}_i \varepsilon_j , \tag{2.5}$$

and $R_{ij}^{\min}$ values are based on the arithmetic mean between $R_i^{\min}$ and $R_j^{\min}$,

$$R_{ij}^{\min} = \frac{\left(R_i^{\min} + R_j^{\min}\right)}{2}. \tag{2.6}$$

Since bonds are modeled with a harmonic potential, bond elongation leads inevitably to an increase in energy and a bond can never be broken. Thus, chemical reactions cannot be modeled with a MM *force field*.

2.2 Semi-empirical Methods

At present, *ab initio* methods, density functional theory (DFT) methods, and semi-empirical (SE) methods serve as major computational tools for quantum chemical simulations. *Ab initio* based simulations can be extremely expensive in terms of the computational resources required. This limits their ability to handle large sized molecules. On the other hand SE molecular orbital methods provide a means of obtaining computational results for large sized organic and inorganic molecules in a fraction of the time as their *ab initio* counterparts. The speed up in computational time is due in part to the explicit consideration of only the valence electrons and representing them with a minimal basis set. A further reduction in computational effort is achieved by neglecting the products of all basis functions on different atoms. The approximations made in order to speed up SE based calculations are compensated for by parameterizing the remaining integrals, of which their values are assigned based on high-level calculations or experimental data.[7]

A number of SE methods currently exist such as Hückel Molecular Orbital (HMO) method, Pariser-Parr-Pople (PPP) method, Extended Hückel Theory (EHT), Complete Neglect of Differential Overlap (CNDO), Intermediate Neglect of Differential Overlap (INDO), Modified Intermediate Neglect of Differential Overlap (MINDO) and Neglect of Diatomic Differential Overlap (NDDO). In addition numerous NDDO type methods are currently available, including, Modified Neglect of Differential Overlap (MNDO), Austin Model 1 (AM1), Parameterized Model 3 (PM3), and Modified Neglect of Differential Overlap with *d*-orbital interactions

(MNDO/d), just to name a few. The background pertaining to these as well as other important SE methods shall be discussed in the sections that follow.

2.2.1 HMO, PPP and EHT

The Hückel Molecular Orbital (HMO) method is the earliest, simplest and most prominent π-electron theory for treating conjugated molecules.[8,9] It was used to predict the properties and reactivities of planar conjugated compounds. A major drawback of the HMO method is that it failed to treat electron repulsion. The first SE π-electron theory that included the effect of electron repulsion between valence electrons and hence improved upon the HMO method is the Pariser-Parr-Pople (PPP) method. Both HMO and PPP methods are only applied to planar conjugated molecules, but PPP allows heteroatoms other than hydrogen. Today the PPP method is still used in cases that require minimal electronic effects. Extended Hückel Theory (EHT) is a molecular orbital theory that takes into account all valence electrons in the molecule and is applicable to non-planar molecules.[10,11] Even though EHT is very poor at predicting molecular geometries it is still used for modeling inorganic compounds and computing band structures in a reasonable CPU time.

2.2.2 CNDO and INDO

All-valence-electron SE methods using the Zero Differential Overlap (ZDO) approximation, such as Complete Neglect of Differential Overlap (CNDO),[12-15] Intermediate Neglect of Differential Overlap (INDO)[12,15] and Neglect of Diatomic Differential Overlap (NDDO)[12,13] were proposed by Pople and his co-workers starting from the mid-1960s. The ZDO approximation greatly simplifies the computation of wave functions by eliminating many of the two-electron integrals. At the ZDO approximation all three- and four-center integrals vanish. The simplest of the all-valence-electron Neglect of Differential Overlap (NDO) models is CNDO. In this model only the outer valence electrons are explicitly treated, the inner-shell electrons are taken as a part of the atomic core.[16] It has proven useful for some hydrocarbon results but little else.[17] Practically all CNDO calculations are performed using the CNDO/2 method,[18] which is an improved parameterization over the original CNDO/1 method.[12] In the INDO approximation, the primary modification to the CNDO approximation is that one-center repulsion integrals between atomic orbitals on the same atom are not neglected. However, the INDO approximation shares

with CNDO the inadequate representation of electron repulsions involving atomic orbitals with directional properties. Today, the INDO method is still used as an initial guess for *ab initio* calculations. In 1973 Zerner and Ridley[19] developed the Zerner INDO method (ZINDO) which is also called spectroscopic INDO (INDO/S). This is a re-parameterization of the INDO method specifically for the purpose of reproducing electronic spectra results. The method is also used for modeling transition metal systems. It predicts ultra-violet (UV) transitions well, with the exception of metals with unpaired electrons.[17] However, it produces generally poor results for geometry optimization. Another INDO based technique is Symmetrically Orthogonalized Intermediate Neglect of Differential Overlap (SINDO).[20-22] It is a method that explicitly takes ortogonalization transformations of the basis functions into account and treats inner orbitals by a local pseudopotential.[9] SINDO appears to perform well but has not found the wide range of acceptance of the NDDO based methods.

2.2.3 NDDO

The Neglect of Diatomic Differential Overlap (NDDO) method was an improvement on INDO, since it neglects differential overlap only when the atomic orbitals are on different atoms. Thus dipole-dipole interactions are retained and expressed in terms of integrals that are calculated either from atomic orbitals or determined empirically.[23] Most modern SE models are NDDO models and a number of them shall be described in the sections that follow.

2.2.4 MINDO

There are three Modified Intermediate Neglect of Differential Overlap (MINDO) methods that were introduced by Dewar and co-workers namely MINDO/1, MINDO/2 and MINDO/3.[24-26] With this method Dewar aimed to calculate ground-state properties, in particular heats of formation (ΔH_f) and molecular geometries, with chemical accuracy, such as bond lengths of 0.1 pm, bond angles of 0.1^o and ΔH_f that are correct to 0.1%.[9,25] The reason for modifying INDO was to remove deficiencies in the analytical calculation of the one-electron repulsion integrals. As such, with MINDO, the integrals are evaluated by using parameters and fitting these parameters to experimental data. Combining MINDO with the Davidson-Fletcher-Powell geometry optimization routine[27] resulted in a parameterization program that was able to accept initial geometries as input and derive the associated minimum energy structures. It can be

said that MINDO represented a very big step toward encouraging chemists to use molecular orbital calculations in the interpretation of experimental data. The third version of MINDO (MINDO/3) was by far the most reliable and was accepted to be the first modern SE method. However, there were a number of limitations to the MINDO method, such as too positive ΔH_f for unsaturated molecules, too large bond lengths and too negative ΔH_f for molecules that contained adjacent atoms with lone pairs.[9]

2.2.5 MNDO

To overcome the limitations of MINDO, Dewar and Thiel introduced the Modified Neglect of Differential Overlap (MNDO) method in 1977.[28] The method evaluates one-center two-electron integrals based on spectroscopic data and evaluates other two-electron integrals using the idea of multipole-multipole interactions from classical electrostatics.[9] Rather than determine various integrals analytically, numerical parameters are adjusted to fit experimental data as in MINDO. MNDO was parameterized to reproduce ΔH_f as well as geometrical properties of stable molecules using ionization potentials and dipole moments. During the parameterization the overlap terms, β_s and β_p, and Slater orbital exponents, ζ_s and ζ_p, for *s*- and *p*-atomic orbitals were fixed. This meant that they were not parameterized separately, instead they were just considered as $\beta_s = \beta_p$ and $\zeta_s = \zeta_p$. These terms are explained in detail in Chapter 3. Despite the advances achieved with MNDO the method does have some disadvantages, such as its inability to model intermolecular systems containing hydrogen bonds accurately when the atoms are separated by a distance within the sum of their van der Waals radii. In addition, molecules possessing hypervalency are considerably unstable, four-membered rings are too stable, rotational barriers are often underestimated, activation barriers are too high, electronic excitation energies are underestimated and conformational preferences are sometimes not reproduced.[29]

2.2.6 AM1

Austin Model 1 (AM1) was introduced by Dewar et al.[30] in 1985 as a modification to and a re-parameterization of the general theoretical model found in MNDO. The major difference is the addition of attractive and repulsive Gaussian Core Functions (GCFs) to the description of the nuclear repulsion term to overcome MNDO's hydrogen bond problem.[9] In addition, instead of

fixing the overlap terms and Slater orbital exponents, as was the case in MNDO, the terms were parameterized separately during the development of AM1. The added Gaussians, overlap terms and Slater orbital exponents significantly increased the number of parameters to be parameterized from 7 per atom (in MNDO) to 13-19 per atom (in AM1).[9] The main gains of AM1 were its ability to reproduce hydrogen bonds and the promise of better estimation of activation energies for reactions.[30] Unfortunately, there are known limitations to AM1, such as:

i) Predicting rotational barriers to be one-third the actual barrier.
ii) Predicting five-membered rings to be too stable.
iii) Predicting hydrogen bonds with the correct strength, but often the wrong orientation.
iv) Geometries of compounds possessing hypervalent atoms are predicted poorly.[17]

Despite the disadvantages, AM1 has been used very widely because of its performance and robustness. The method has retained its popularity and after a few improvements is still used by numerous researchers today.[31-34]

2.2.7 PM3

The parameterization of AM1 was essentially done by hand, taking the one-center two-electron (G_{ss}, G_{sp}, G_{pp}, G_{p2} and H_{sp}) parameters from atomic data and varying the rest until a satisfactory fit had been obtained. Since the optimization was done by hand, only a few reference compounds could be included.[35] In 1989 Stewart developed Parametric Method Number 3 (PM3)[36] in which the optimization of parameters was a completely automated process. This was done by deriving and implementing formulas for the derivative of a suitable error function with respect to the parameters. All parameters could then be optimized simultaneously, including the one-center two-electron terms, and a significantly larger training set with several hundred data could be employed.[35] The optimization process does however, still require some human intervention in selecting the experimental data and assigning appropriate weight factors to each set of data. PM3 also differs from AM1 in the number of Gaussian terms used in the nuclear (core-core) repulsion function. Instead of using up to four Gaussians per atom as in AM1, PM3 makes use of only two Gaussians per atom. Although based on AM1, PM3 did not enjoy Dewar's blessing. The reason for this is due to the fact that Dewar felt that PM3 represented at best an only marginal improvement over AM1 and that a new SE method should make previous

ones essentially obsolete.[37] Stewart defended his approach by stating that if PM3 was only a marginal improvement over AM1, then AM1 was only a marginal improvement over MNDO.[38] Dewar also objected strongly to any proliferation of computational chemistry methods, whether it be in the realm of *ab initio* basis sets or of SE methods.[3]

Nonetheless, overall PM3 does predict heats of formation that are more accurate than MNDO or AM1. Hypervalent molecules are predicted more accurately. Hydrogen bond angles are more accurate than those of AM1. However, as with AM1, there are known problems with PM3, these include:

i) The rotational barrier of the amide bond is much too low and in some cases almost non-existent.
ii) Hydrogen bond energies are not as accurate as those of AM1.
iii) There is a tendency to predict an sp^3 nitrogen as always being pyramidal.
iv) Bonds between Si and halide atoms are too short.
v) Predicts incorrect electronic states for germanium compounds.

Despite all of this AM1 and PM3 are still widely used, with AM1 being the most popular.

2.2.8 MNDO/d

For NDDO methods discussed thus far, such as MNDO, AM1 and PM3 only *sp*-basis sets are used and no *d*-orbitals are included in their original implementation. Hence they cannot be applied to transition metal compounds. MNDO and AM1 were not designed to treat hypervalent compounds, but in the parameterization of PM3 considerable effort was made to overcome such deficiencies. In addition, *ab initio* calculations have shown that *d*-orbitals are significant for quantitative accuracy in the hypervalent compounds of main group elements.[9] Because of these limitations and deficiencies, it was necessary to extend the MNDO formalism to include *d*-orbitals. In 1992, Thiel and Voityuk introduced MNDO/d, the first NDDO model to include *d*-orbitals.[39,40] MNDO/d explicitly contains *d*-orbitals for heavier atoms starting from the second row in the periodic table, but it uses the theory and parameters of the original MNDO method for elements hydrogen–fluorine. In MNDO/d two-center two-electron integrals are calculated using the original point-charge[41] model which was used in MNDO, AM1 and PM3. The integrals are expanded in terms of SE multipole-multipole interactions. For an *spd*-basis, there are 45 distinct

one-center charge distributions that are associated with multipoles up to hexadecapoles. In the case of MNDO/d all monopoles, dipoles and quadropoles of these charge distributions are included whereas all higher multipoles are neglected.[39] MNDO/d predicts the point groups and ΔH_f (producing a smaller mean absolute error) of hypervalent compounds more accurately compared to MNDO, AM1 and PM3.[42]

2.2.9 SAM1

Semi *Ab Initio* Method 1 (SAM1) was the last SE method to be reported by the Dewar group.[43] It is not a straightforward extension of the NDDO formalism, but represents a rather different approach to constructing the Fock matrix.[44] In SAM1, the two-electron repulsion integrals are *ab initio* integrals that are evaluated from contracted Gaussian basis functions (STO–3G) fit to Slater-type orbitals using standard methods.[45] The method uses a parameterization to estimate the correlation effects. For organic molecules too large for correlated *ab initio* calculations, this is a reasonable way to incorporate correlation effects.[17] SAM1 has parameters for H, Li, C, N, O, F, Si, P, S, Cl, Fe, Cu, Br and I. As with MNDO/d, SAM1 includes *d*-orbitals for heavy atoms starting from the second row of the periodic table. The method is unfortunately only available within the commercial software package, AMPAC 9.[46]

2.2.10 PM3(tm) and AM1/d

The approach used during the development of MNDO/d was adopted by researchers to extend Hamiltonians such as AM1 and PM3 to include *d*-orbitals. Hehre and co-workers added *d*-orbitals to the original PM3 Hamiltonian and they named this method PM3(tm), where 'tm' emphasizes a focus on transition metals.[47] During the parameterization process used for PM3(tm) only geometrical data (primarily from X-ray crystallography) was included in the fitness function. Properties such as energies, dipole moments and ionization potentials were not taken into account. Thus, PM3(tm) may be regarded as an efficient way to generate reasonable molecular geometries whose energies may then be evaluated using more reliable levels of theory.[44] The method is currently only available in the commercial software package SPARTAN 8.0 and above.[48] A few years after the development of PM3(tm), Voityuk and Rösch extended AM1 to include *d*-orbitals for molybdenum, which they appropriately named AM1/d.[49] In

addition to the *d*-orbital inclusion, AM1/d has a modified core-core repulsion. The Gaussian-type functions, used in the original AM1 Hamiltonian to refine the core-core repulsion term, have been excluded and two bond specific parameters have been included in its place. Another variation of AM1/d was introduced in 2003 by Lopez and York,[31] where the *d*-orbital inclusion departed from the MNDO/d formalism and the Gaussian core-core terms were retained, rather than excluded as in the Voityuk approach.[49] With this method the authors obtained parameters for phosphorus which were used to treat nucleophilic attacks of biological phosphates. The results obtained demonstrated that the strategy of developing SE parameters specific for biological reactions offers considerable promise for application to large-scale biological problems.

2.2.11 PM3-PIF and PM3-MAIS

Despite the noticeable improvement provided by PM3 over MNDO by addition of GCFs to the core-core repulsion interactions, numerous researchers[50-54] have pointed out some serious weaknesses in the method. Their investigations proved that:

i) The use of GCFs in the core-core repulsion interactions is not sufficient enough to ensure a good estimation of the intermolecular interaction energy.
ii) The use of GCFs introduces spurious artifacts into the potential energy surface.

In order to address the problems Bernal-Uruchurtu et al.[55] developed a method in which the GCF was replaced with a simple function exhibiting the correct physical behavior in the whole range of intermolecular separation distances. The method, entitled PM3-parameterized interaction function (PM3-PIF), introduces a sum of atom-pair contributions (similar to those in molecular mechanics models), each one having five adjustable parameters. This method exhibited some valuable features, such as:

i) It has the correct physical behavior of a function that is intended to fit an intermolecular potential energy surface.
ii) It is flexible enough, but contains a limited number of adjustable parameters, comparable to that employed in the core-core repulsion function of SE theories.
iii) The additional computational cost is negligible.

Despite these features PM3-PIF does require that intermolecular and intramolecular terms be treated separately. In order to overcome this drawback the authors developed a new PM3 core-core correction function that behaves like the original PM3 term at short interatomic distance and goes to the PIF function as distances increase.[56] This function was entitled the method adapted for intermolecular studies (MAIS), giving rise to PM3-MAIS, which reproduced proton transfer barriers in very good agreement with most refined *ab initio* methods.

2.2.12 PDDG/MNDO and PDDG/PM3

In 2002 Repasky et al.[57] developed a method in which they added functional group information into the core-core repulsion term of standard MNDO and PM3 Hamiltonians. A number of considerations in designing the function had to be taken into account, these included:

i) The interactions introduced by the function must make small contributions to the overall molecular energy, or they may overwhelm the electronic portion and adversely alter optimized molecular geometries.
ii) The individual terms must be able to differentiate between a wide range of functional groups based on molecular geometries with a limited number of parameters.
iii) Bond specific parameters must not be used to avoid the trap of an exponentially expanding parameter set.
iv) No parameters must be introduced for specific functional groups or interactions.
v) The corrections must not introduce significant errors in molecules with nonstandard bonding, such as charged species and transition states.

The most successful function, which fulfilled all of the criteria mentioned above, was composed of four weighted Gaussians for heterodimer atom pairs and three weighted Gaussians for homodimers. In addition the authors reevaluated the procedure for deriving EISOL parameters, which are dependent, in a nonsystematic way, on the values of all one-center parameters within the SE formalism. Finally a re-parameterization of the standard MNDO and PM3 parameters was carried out for H, C, N, O, F, Cl, Br, I, S, Si and P.[57-59] This gave rise to the Pairwise Distance Directed Gaussian (PDDG) methods, PDDG/MNDO and PDDG/PM3.

2.2.13 AM1*

In 2003 Winget et al. developed AM1*,[32] which is an extension of the original AM1 SE molecular orbital technique. AM1* has been based on AM1, rather than MNDO or PM3, because AM1 reproduces the energies of hydrogen bonds (but not their geometries) relatively well and generally performs better for rotation barriers of partial double bonds (such as the C–N bond in amides) than the other two methods.[32] AM1* uses the AM1 parameters and theory unchanged for the elements H, C, N, O and F. For all other elements (P, S, Cl, Al, Si, Ti, Zr, Cu, Zn, Br, I, V, Cr, Co, Ni, Mn, Fe, Pd and Ag)[32,34,60-65] an additional set of *d*-orbitals were included in the basis set and a modified core-core repulsion function was utilized. This shall be discussed in more detail in Chapter 3. The use of original AM1 parameterization elements limits AM1*'s ultimate accuracy in some cases.[9] However, results obtained with the AM1* Hamiltonian have shown that the method performs very well compared to other methods, such as MNDO/d, PM5 and PM6.[63]

2.2.14 PM3CARB-1

The performance of PM3 was evaluated by McNamara et al.[66] in 2004, where they applied it to a number of carbohydrate systems. It was found that for a given anomer of *β*-glucopyranose the PM3 Hamiltonian predicted a $^1C_4$ ring conformation as more stable than $^4C_1$, contrary to high level QM calculations and experiment.[67,68] The authors realized that the only means to correct for this erroneous prediction was to re-parameterize the SE method for specific bonding situations. As such, they re-parameterized the PM3 Hamiltonian in a fashion analogous to fitting of a classical force field, basing their strategy on small molecule carbohydrate analogues. The strategy adopted during the parameterization followed the specific reaction parameter (SRP) approach of Rossi and Truhlar,[69] whereby selected parameters of a SE MO method are adjusted to fit *ab initio* data for a specific reaction. The resulting parameters produced PM3CARB-1, which was in general more able to accurately predict structures and energetics of a set of small carbohydrate analogues as compared to the standard PM3 Hamiltonian. In addition the $^1C_4$ conformers of *β*-glucopyranose are correctly ranked by PM3CARB-1 as being less favorable than $^4C_1$ conformers. However, despite the improvements achieved with PM3CARB-1, the method does require further development before applying it more generally to carbohydrate modeling.[66]

2.2.15 RM1

In 2006, Rocha et al. developed a method called Recife Model 1 (RM1).[70] The method is a re-parameterization of AM1 were properties such as, ΔH_f, dipole moments, ionization potentials and geometric variables (bond lengths and angles) are used in the parameterization procedure. Unlike AM1, and similar to PM3, all RM1 parameters have been optimized. No changes were made to the original AM1 formalism or to the approximations used in AM1. RM1 has been re-parameterized for ten elements (H, C, N, O, F, P, S, Cl, Br and I) and for organic molecules the method has shown increased accuracy when compared to other NDDO methods.[70]

2.2.16 AM1/d-PhoT

In 2007, Nam et al.[33] wished to model phosphoryl transfer reactions with a SE based method (AM1). They realized however, that AM1 with its original set of approximations and parameters would be inadequate to accurately model the types of systems they were interested in. As such the authors conducted a re-parameterization of AM1, focusing specifically on atoms hydrogen, oxygen and phosphorus. Prior to carrying out the re-parameterization the authors ensured that an *spd* basis was applied to the phosphorus atom, in order to provide a better description for the hypervalent nature of the atom. Furthermore the authors had to consider both the positive and negative aspects of using the standard AM1 Hamiltonian, which are:

i) AM1 was initially developed to offer improvement for hydrogen bonding relative to MNDO.
ii) AM1 has the problem that it over-stabilizes hypervalent structures because of the artificially attractive core-core interactions.

With these points in mind AM1/d-PhoT was designed to keep the core-core interactions for hydrogen bonding, but to turn these interactions off for phosphorus bonding where the *d*-orbitals allow proper hybridization and accurate representation of hypervalent species. Toward this end, a scale factor was introduced into the Gaussian core-core terms of the AM1 Hamiltonian. This scale factor was allowed to vary from zero to one (values of 0 recover the conventional MNDO core-core model, whereas values of 1 recover the AM1 core-core model).[33] The resulting AM1/d-PhoT parameters were tested in the gas phase and in solution using a hybrid quantum mechanics/molecular mechanics (QM/MM) potential. The results obtained indicate that the

method provides significantly higher accuracy than MNDO, AM1 and PM3 methods. Moreover for the transphosphorylation reactions studied, AM1/d-PhoT was in close agreement with the density functional calculations carried out at the B3LYP/6-311++G(3df,2p) level.[33]

2.2.17 PM6

In close proximity to the release of AM1/d-PhoT, Stewart developed Parametric Method Number 6 (PM6),[71] which is a method parameterized for 70 elements of the periodic table. This was achieved by making several changes to the NDDO core-core interaction terms, utilizing a different parameter optimization methodology and inclusion of *d*-orbitals to the basis set. The inclusion of *d*-orbitals was used to enhance the treatment of main group and transition metal systems. Stewart believed that there were three sources of error in SE methods,[71] namely:

i) Reference data may be inaccurate or inadequate.
ii) The set of approximations may include unrealistic assumptions or may be too inflexible.
iii) The parameter optimization process may be incomplete.

PM6 was developed as a way to circumvent the above mentioned sources of error. As a result the method produced an average unsigned error (AUE) of 8.0 kcal/mol between calculated and reference heats of formation for 4492 species. For a subset of 1373 compounds involving only H, C, N, O, F, P, S, Cl and Br the AUE for PM6 was 4.4 kcal/mol.[71] The equivalent errors for RM1, PM3 and AM1 were 5.0, 6.3 and 10.0 kcal/mol, respectively. The PM6 Hamiltonian is freely available to academics in the software package MOPAC2009.[72]

2.2.18 OMx

It is generally accepted that proper orthogonality of orbitals is essential to account for the dominant contributions to the barriers that arise from Pauli exchange repulsion.[73] Due to the approximation of ZDO, these orthogonalization effects are neglected in some established SE methods, such as MNDO, AM1 and PM3. The OMx methods include orthogonalization corrections into the one-electron part of the Hamiltonian from the transformation of the secular equations from a non-orthogonal to an orthogonal basis. These corrections are incorporated only into the one-center one-electron terms in orthogonalization model 1 (OM1),[74,75] but also into the two-center one-electron terms in OM2.[73,76] Some of the latter tend to be small and disregarded in

OM3,[77] which is thus a simplified (and somewhat faster) variant of OM2. The results from the OMx methods are generally superior to those from MNDO, AM1 and PM3.[78]

2.2.19 PM3^{MS}

A few years after the introduction of PM3CARB-1, Mane and Klobukowski[79] developed PM3^{MS} (PM3 monosaccharide) which was also a re-parameterization of the standard PM3 Hamiltonian. The parameterization was conducted in a similar manner as that of PM3CARB-1 in which the specific reaction parameter approach of Rossi and Truhlar[69] was utilized. However, instead of adjusting parameters in order to fit high-level QM data for a specific reaction, as was the case for PM3CARB-1, with PM3^{MS} the authors adjusted selected parameters to fit energies and geometries of eight conformers of D-galactopyranose, eight conformers of D-glucopyranose, and six conformers of D-mannopyranose. The method represented a major improvement over the original PM3, correctly predicting the energies of the lowest and highest energy conformers present in the PM3^{MS} training set. In addition, relative energies of monosaccharides outside of the training set outperformed results obtained with the original PM3. Despite the improvements, PM3^{MS} does produce reference molecule geometries that are overestimated by more than 0.10 Å. The parameterization was also not very extensive and a more thorough optimization could provide parameters that are superior to the current PM3^{MS} set.[79]

2.2.20 PM7

During the course of 2012 Stewart introduced a new method called Parametric Method Number 7 (PM7).[80] The method was developed to tackle known faults with its predecessor, PM6, such as:

i) Missing repulsion between Na–Na, Br–N, Br–O, Br–Br, S–N, S–S, S–O, S–Cl, I–N, I–O and I–I pairs.
ii) Production of an infinite error when the method was applied to crystal structures.
iii) The incorrect prediction that a Si–O–H system is linear.
iv) Procedural faults not detected during the development of the method.
v) Poor description of dispersion and hydrogen bond interactions.
vi) Low accuracy in reproducing barrier heights for reactions.

PM7 showed significant improvement over its predecessor (PM6) especially as far as geometries are concerned.[80] Along with PM7 came the development of PM7-TS, which can be used to describe chemical reactions efficiently by being able to model activation barrier heights. Both methods can be found in the software package MOPAC2012.[81]

2.2.21 Dispersion and Hydrogen bonding

The SE methods discussed above are incapable of modeling dispersion bound complexes because the form of the SE wave function completely neglects electron correlation. Even quantitatively modeling dispersion-bound macromolecular systems, such as complexes of carbon nanostructures, is out of the question, since SE methods predict such complexes to be unbound.[82] This accuracy can be dramatically improved by adding an empirical correction, comprised of an explicit R^{-6} term, which is used to describe interatomic dispersive interactions. McNamara and Hillier[83] used this methodology, together with optimization of 18 parameters of AM1 and PM3, to add an empirical correction term to these methods. These methods were referred to as AM1-D and PM3-D, respectively. The methods accurately predict intermolecular interaction energies, but lack the ability to reproduce heats of formations.[82]

In 2009 Řezáč et al.[84] introduced an empirically corrected PM6 method augmented with dispersion and hydrogen bonding corrections (DH), entitled PM6-DH. The dispersion correction was based on work proposed by Jurečka and co-workers,[85] whereas an improvement for H-bonding was achieved by including a second correction term involving three parameters. This term was particularly simple, depending only on interatomic distance, an angle and the partial charges on the hydrogen and acceptor atoms (oxygen or nitrogen) involved. A dramatic improvement was followed shortly by DH2,[86] in which the hydrogen bond energy term was extended to include some torsion angles. However, two problems exist in the DH2 correction:

i) The derivative of the charge with respect to the coordinates, which is expensive to calculate, enters the expression for the gradient of the correction as zero. This approximation cannot be used in certain cases, such as in accurate optimizations or in molecular dynamics.

ii) A proton transfer along a hydrogen bond exhibits a discontinuous potential energy surface.

In order to address the aforementioned problems, Korth[87] introduced a new correctional term in 2010 entitled DH_+. This method depended only on the geometry of the system, not on partial charges, thus ensuring compliance with the variational principle.

Recently, Řezáč and Hobza[88] identified a few problems with DH_+, such as:

i) The earlier version of the H-bonding correction used the atomic charges and thus naturally described strong H-bonds involving charged groups. In DH_+, the same parameters are used for neutral and charged H-bonds, which leads to an underestimation of the interaction in charged systems.

ii) The linear terms in both DH2 and DH_+ do not have smooth first derivatives, which makes it impossible to optimize the geometry of some systems.

To address these problems Řezáč and Hobza[88] developed D3H4. With this method the authors wished to preserve the improvements brought by DH_+, solve its poor performance in charged systems and, importantly, simplify the form of the correction. Unlike its predecessors, D3H4 has a smooth potential energy surface as well as first and second derivatives. Another significant feature is that the correction potential is strictly local and does not have to be evaluated for more distant potential H-bonds. As a result the computational expense grows only linearly with the size of the calculated system. To date the DH correction has been parameterized for PM6; DH2 possess parameters for PM6, AM1, SCC-DFTB (discussed in section 2.3) and OM3; DH_+ was parameterized for PM6, AM1, SCC-DFTB, OM3 and a number of force field methods and finally D3H4 can be utilized with PM6, RM1, OM3, PM3, AM1 and SCC-DFTB methods.

Other variants of dispersion and hydrogen bond correctioned SE methods do exist, such as AM1-FS1,[82] but shall not be discussed further in this work.

The mathematical detail pertaining to all of the correctional terms mentioned above shall be discussed in Chapter 3.

2.3 SCC–DFTB

Self-consistent charge density functional tight-binding (SCC–DFTB) is an approximate method that is derived from density functional theory (DFT) by neglect, approximation and parameterization of interaction integrals.[89,90] SCC–DFTB constitutes an alternative to the traditional SE methods mentioned above. However, it is *not* a SE method in the strict sense,

since its parameterization procedure is completely based on DFT calculations, no fit to empirical data has to be performed.[90] In addition SCC–DFTB is a non-orthogonal method, i.e., it is based on a non-orthogonal basis set (SE methods have also been extended to non-orthogonality in the OMx methods mentioned above).[74-76] In the framework of tight-binding theory, this has been emphasized to be a key factor for transferability.[91] Transferability denotes the ability of a parameterized method to perform sufficiently well for chemical based environments not included in the parameterization procedure.

2.4 Hybrid QM/MM methods

An interest in understanding solvent structure, nanostructured materials, condensed-phase reactions, catalytic systems, including designer zeolites and enzymes, and modeling systems over longer time scales that reveal new mechanistic details represents some examples of a situation that requires the explicit representation of a large system.[44] For reasons of efficiency, such representation is typically carried out at the MM level. However, these methods are classical by definition and do not describe quantum effects, such as processes involving bond-making and bond-breaking, i.e. chemical reactions, which are essential in the overwhelming majority of problems. To model such processes adequately, quantum mechanical (QM) methods are required. These considerations lead directly to the idea of separating a molecular system into two (or more) regions in such a way so as to find the quantum effects taking place overwhelmingly in only one of them, while the other region/s are considered by classical methods.[92] Hybrid quantum mechanics/molecular mechanics (QM/MM) methods are based on this idea.

The seminal contribution by Warshel and Levitt[93] in 1976 marked the beginning of the QM/MM era. They introduced the QM/MM concept, presented a method with many of the features that are now considered essential in the field and applied it to an enzymatic reaction.[94] However, the QM/MM approach only found widespread acceptance in the 1990s when Field et al.[95] described, in detail, the coupling of SE QM methods to the CHARMM force field and carefully evaluated the accuracy and effectiveness of the QM/MM treatment against *ab initio* methods and experimental data. The QM/MM approach has been established as a valuable tool not only for the modeling of biomolecular systems, but also for the investigation of inorganic/organometallic[96,97] and solid state systems[98,99] and for studying processes in explicit solvent.[100-102]

A QM/MM method (Figure 2.2) treats a localized region, e.g. the active site, and its neighbors in an enzyme with QM methods and includes the influence of the surroundings, e.g. the protein environment, with an MM force field.[103] The QM/MM energy is modeled as the sum of the QM energy, the MM energy and a QM/MM interaction term,

$$E_{tot} = E_{QM} + E_{MM} + E_{QM/MM} . \tag{2.7}$$

When using a molecular orbital description for the quantum mechanical region, the QM energy (based on the SE methodology) can be calculated as outlined in Chapter 3. The MM energy is determined as in Section 2.1.

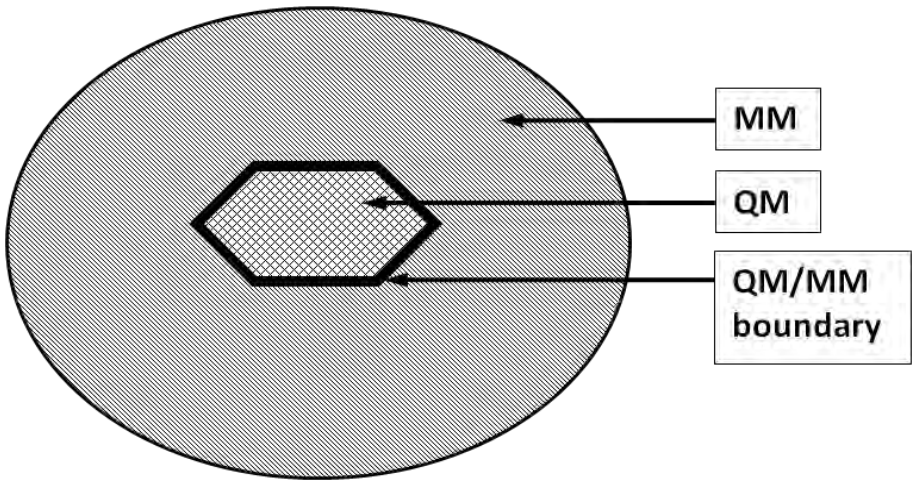


Figure 2.2: Partitioning of a QM/MM system.

The QM/MM coupling term, $E_{QM/MM}$, defines a particular QM/MM method. In accordance with the interactions considered in the force field (eq. 2.1), it includes bonded, non-bonded (van der Waals) and electrostatic interactions between QM and MM atoms,

$$E_{QM/MM} = E^{b}_{QM/MM} + E^{vdW}_{QM/MM} + E^{el}_{QM/MM} . \tag{2.8}$$

The sections that follow provide more detail related to the individual terms which contribute to $E_{QM/MM}$. The electrostatic coupling term (Section 2.4.1) is normally the most important and also the most technically involved one. The van der Waals interaction and bonded terms are discussed

in Section 2.4.2. Finally, the various ways that have been devised to treat covalent bonds across the QM/MM boundary are presented in Section 2.4.3.

2.4.1 The electrostatic QM/MM interaction

The electrostatic coupling between the QM charge density and the charge model used in the MM region can be handled at numerous different levels of sophistication, characterized essentially by the extent of mutual polarization and classified accordingly as mechanical embedding (model A), electrostatic embedding (model B) and polarized embedding (model C and D).[94,104]

In the case of mechanical embedding, the QM/MM electrostatics is treated at the MM level. The charge model of the MM method (typically rigid atomic point charges but other approaches, e.g. bond dipoles, are also possible) is simply applied to the QM region as well. Both the QM and MM region are unpolarized in this case and the QM charge density comes from a gas-phase calculation (without MM environment).[105] This often has drawbacks, such as:

i) The treatment requires an accurate set of MM parameters, such as atom-centered point charges for both the QM and MM regions. It is relatively easy to get such parameters for the MM region, but the problem lies in getting such parameters for the QM region, where reactions are taking place, since this was the main reason for shifting from MM to QM in the first place.

ii) The potential perturbation of the electronic structure of the QM region due to the electrostatic interaction between the QM and MM is ignored, which results in atom-centered charges in the MM region polarizing the QM region and altering its charge distribution. This is especially problematic if the reaction taking place in the QM region is accompanied by charge transfer.

iii) For systems having several electronic states (e.g. an open-shell system containing transition metals) close in energy, the polarization could change the energetic order of these states. This results in prediction of different ground states with different charge and/or spin distributions.[103]

These drawbacks result in an electrostatic treatment that will often not be accurate enough, especially in the case of very polar environments (as in most biomolecules).

The major shortcomings of mechanical embedding can be eliminated or avoided by performing the QM calculation in the presence of the MM charge model. This can be done by incorporating the MM point charges as one-electron terms in the QM Hamiltonian,

$$H_{QM/MM}^{el} = \sum_{\alpha \varepsilon QM}^{M} \sum_{J \varepsilon MM}^{L} \frac{q_J Q_\alpha}{|R_\alpha - R_J|} - \sum_{i}^{N} \sum_{J \varepsilon MM}^{L} \frac{q_J}{|r_i - R_J|}, \quad (2.9)$$

where q_J is the MM point charges located at R_J, Q_α is the nuclear charges of the QM atoms at $R\alpha$, r_i designates the electron position and indices i, J and α run over the N electrons, L point charges and M QM nuclei, respectively.

In such a scheme, known as electrostatic embedding, the electronic structure of the QM region can adapt to changes in the charge distribution of the environment and is automatically polarized by it. The QM/MM electrostatic interaction is treated at the QM level, thereby providing a more advanced and more accurate description than a mechanical embedding scheme. Electrostatic embedding does, however, increase the computational requirements, especially for the calculation of the Coulomb forces as a result of the QM density acting on the (many) MM point charges.[94] Special care is required at the QM/MM boundary, where the MM charges are placed in the immediate proximity to the QM electron density and can lead to overpolarization. This problem is more pronounced when the boundary runs through a covalent bond. The detail related to the treatment of such a boundary shall be provided in Section 2.4.3.

As electrostatic embedding accounts for the interaction of the polarizable QM density with rigid MM charges, the next step is to introduce a flexible MM charge model that is polarized by the QM charge distribution. These so called polarized embedding schemes can be further divided into approaches where the polarizable charge model in the MM region is polarized by the QM electric field, but does not itself act back on the QM density (model C); and fully self-consistent formulations that include the polarizable MM model into the QM Hamiltonian and therefore allow for mutual polarization (model D).[94,104] There are various models for treating polarization in classical simulations.[106-109] There are, however, no established polarizable biomolecular force fields. To date there are a limited number of polarized embedding QM/MM simulations which have been utilized, with majority of the simulations being restricted to explicit solvation (in particular, hydration), where the solute is treated at the QM level and the solvent by a polarizable force field.[110-116] The application of polarizable embedding is expected

to become more popular when polarizable biomolecular force fields are better established and used more often as MM components in QM/MM work.

2.4.2 Non-bonded and bonded QM/MM interactions

The non-bonded (van der Waals) and bonded contributions to the QM/MM coupling term, given in eq. 2.8, are considerably simpler than the electrostatic treatment described above, as they are handled purely at the MM level. Non-bonded interactions are described by a Lennard-Jones (LJ) potential (eq. 2.4), which implies that all interactions are calculated at the MM level and therefore, rely on the availability of MM parameters for the atoms present in the QM region. Even if suitable LJ parameters exist for a given configuration, QM atoms can change their character, for example, during the course of a reaction. The question then arises, should MM parameters be switched, say, from a "reactant description" to a "product description" somewhere along the reaction path? Switching between different sets of parameters along a reaction path is not convenient and avoiding this was one of the reasons for moving from MM to QM. Moreover, even if the switching could be done, one does not know at which point along the reaction path it should be done and how suddenly if the change is gradual.[103] In practice however, these complications are alleviated by the short-range nature of the van der Waals (vdW) interaction. While every atom of the QM region is involved in vdW interactions with all the atoms of the MM region, only those closest to the boundary contribute significantly. In principle, the use of a larger QM region pushes the boundary away from the reaction center and helps to alleviate the uncertainty due to parameter choices, but at a price of increasing computational cost.

The formal reservations against using standard MM parameters to describe QM/MM interactions also apply to the bonded (bond stretching, angle bending, torsional, etc.) interactions. The solution to this problem is entirely pragmatic: usually the standard MM parameter set is retained and is complemented as necessary with additional bonded terms not covered by the default assignment rules of the force field.[94]

2.4.3 Covalent bonds that cross the QM/MM Boundary

A critical issue underlying the accuracy and applicability of the combined QM/MM methods for studying enzyme reactions is how to describe the QM/MM boundary across covalent bonds.[95,117-120] The simplest solution to the problem is to circumvent the cutting of

covalent bonds altogether by defining subsystems such that the boundary does not pass through a covalent bond. This can be fulfilled for explicit solvation studies, where the solute is normally described at the QM level, surrounded by MM solvent molecules.[94] Such a favorable situation is sometimes encountered for biomolecular systems; for instance, if an enzymatic reaction involves only partners (substrates, co-factors) that are not covalently bound to the enzyme. Often, however, it is unavoidable that the QM/MM boundary cuts through a covalent bond. In this case, the QM and MM regions must be linked such that the QM region can be treated as a closed-shell system while maintaining the overall structural integrity of the system. Several techniques have been reported for linking the QM and MM regions and for the remainder of this section some of the techniques shall be discussed in more detail.

The link-atom approach is the most straightforward prescription to the boundary atom problem.[95,117,121-123] In this approach link atoms, which are generally hydrogen atoms, are added to saturate the valency of the QM region so as to form a closed-shell system. The link-atom is positioned on the bond that is cut by the QM/MM boundary (Figure 2.3). Classical terms are added so as to hold the QM region in place relative to the MM region. One drawback of the link-atom approach is the introduction of additional degrees of freedom into the system, which complicates the expression of the energy and force, the geometry optimization and molecular dynamics simulation.[124] Although approaches have been made to alleviate these complications within the link-atom framework,[123,125] there is a great deal of interest in the search for approaches without introducing additional atoms into the system.[124]

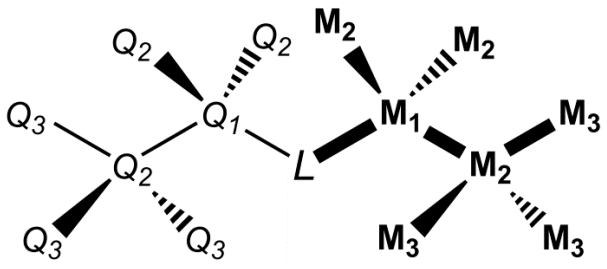


Figure 2.3: Atom labeling at the boundary between QM and MM regions. The QM and MM atoms directly connected are designated Q_1 and M_1, respectively. The first shell of MM atoms (those directly bonded to M_1) is labeled M_2. The next shell, separated from M_1 by two bonds is labeled M_3; and so on. The same naming procedure applies to the QM side; atoms Q_2 are one

bond away from Q_1, Q_3 two bonds away, etc. The link-atom (L) saturates the dangling bond of Q_1.

An alternative to the link-atom method is provided by frozen-orbital approaches in which the closed shell at the boundary QM atom is maintained by using strictly localized orbitals. This class of methods includes:

i) Local Self Consistent Field (LSCF):[126,127] In the LSCF method, developed by Rivail and co-workers, the strictly localized bond orbitals (SLBOs), which are obtained by separate quantum mechanical calculations of small model compounds, are assumed to be transferable for use in proteins. In a QM/MM calculation it is excluded from the SCF optimization and does not mix with other orbitals. It is oriented along the Q_1–M_1 vector and can be described as a sort of frozen lone pair on Q_1 pointing towards M_1 (Figure 2.4a).[94] Numerous studies indicate that the LSCF method can yield good results in energy minimization of reaction pathways in proteins and the assumption of transferability of bond orbitals appears to be valid.[126-129] Although the LSCF method does *not* require the addition of link atoms into the system, the parameters for the localized bond orbitals have to be determined from model studies for each new system in the LSCF treatment.

ii) Frozen Orbitals:[130-132] A variant of the LSCF procedure is the frozen orbital method that differs in some technical details from the original one.[131] In addition, there is a major conceptual difference as compared to most other QM/MM schemes in that the QM/MM interactions at the boundary are heavily parameterized: (a) Several electrostatic correction terms are included that reduce the short-range electrostatic interactions at the interface, following the spirit of 1–2, 1–3 and 1–4 electrostatic exclusion and scaling rules used in force fields. (b) The van der Waals parameters of the QM atoms are re-optimized. (c) Certain classes of hydrogen bonds across the boundary are described by an additional repulsive term. (d) The QM/MM bonded terms are re-optimized, rather than taken directly from the force field.[104] The goal of parameterization is to reproduce, as closely as possible, the conformational and reaction energetics in the boundary region.

iii) Generalized Hybrid Orbital (GHO):[120,133-135] The GHO method is closely related to the LSCF and frozen-orbital approaches in that it constructs localized hybrid orbitals and freezes

some of them. Unlike the LSCF method, the GHO method places a set of localized hybrid orbitals on M_1 instead of Q_1 (Figure 2.4b). M_1 thus becomes a boundary atom, blurring the classification of boundary methods into boundary-atom and frozen-orbital schemes. The orbital pointing towards Q_1 is active and participates in the SCF iterations, while the remaining "auxiliary" hybrids are kept frozen and are allowed to mix with other orbitals.[94] Consequently, the chemical bond connecting the QM and MM fragments is explicitly treated without introducing spurious "link-atoms". Moreover, in contrast to the LSCF approach, the GHO method does not need to be re-parameterized every time a new system is studied.

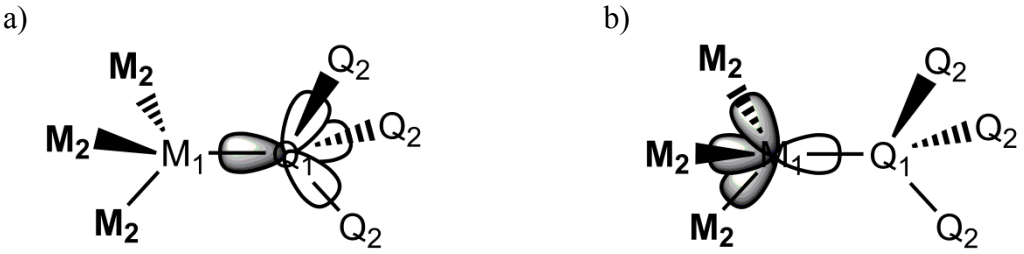


Figure 2.4: Frozen-orbital boundary methods. a) The LSCF method (left) in which a set of localized orbitals is placed on Q_1, one of which (shaded) is kept frozen and points toward M_1. b) The GHO method (right) in which a set of localized orbitals is placed on M_1, one of which (open) is active and points toward Q_1.

2.5 Molecular Dynamics

Although the QM/MM methods mentioned above provide a means of modeling considerably large systems, a major drawback is that they are only capable of modeling the static nature of a given system. When studying biological systems, such as enzymes for example, it is important to simulate the motion of the enzyme as it changes shape on binding to a substrate, rather than just consider a static snapshot of the molecule, which results in a significant loss of detail. As such, in some instances, the QM/MM energy and forces are used with the molecular dynamics (MD) scheme. The objective of MD simulations is the comprehensive sampling of a system's phase space to calculate statistical thermodynamical ensemble averages.[94] Examples include free-energy differences such as reaction, activation or solvation free energies. However, phase space sampling or equilibration of the conformational ensemble is one of the most central problems in all molecular simulations possessing two main obstacles:

i) Every usable potential energy function for molecules must make severe approximations.

ii) Finite computational resources limit the duration of simulations.

Unfortunately, since more accurate theoretical descriptions of molecules tend to be computationally more expensive, trying to overcome one of these obstacles just serves to make the other one more challenging.[136] Therefore, selecting and pre-equilibrating a good starting structure, utilizing enhanced sampling techniques and critically questioning the convergence of results are crucially important for QM MD and QM/MM MD simulations.

The mathematical background pertaining to MD is not presented in this thesis and one is referred to various textbooks[44,137-139] for more detail related to this methodology.

2.6 References

(1) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; ACS Monograph 177, American Chemical Society, Washington, DC, 1982.
(2) Rappe, A. K.; Casewit, C. L. *Molecular mechanics across chemistry*; University Science Books, Sausalito, CA, 1997.
(3) Lewars, E. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*; 2nd ed.; Kluwer Academic Publishers, 2003.
(4) Schlick, T. *Molecular Modeling and simulation: An interdisciplinary guide*; Springer, New York, 2002.
(5) Stewart, J. J. P. *J. Mol. Model.* **2004**, *10*, 155.
(6) MacKerell Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
(7) Lipkowitz, K. B.; Larter, R.; Cundari, T. R.; Boyd, D. B. *Reviews in Compuational Chemistry*; John Wiley & Sons, 2003; Vol. 19.
(8) Hückel, E. *Z. Phys.* **1931**, *70*, 204.
(9) Kayi, H. PhD Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.
(10) Hoffmann, R. *Chem. Phys.* **1963**, *39*, 1397.
(11) Hoffmann, R. *Chem. Phys.* **1964**, *40*, 2445.
(12) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Theory*; McCraw-Hill, Inc.: New York, 1970.
(13) Pople, J. A.; Santry, D. P.; Segal, G. A. *J. Chem. Phys.* **1965**, *43*, S129.
(14) Pople, J. A.; Segal, G. A. *J. Chem. Phys.* **1966**, *44*, 3289.
(15) Pople, J. A.; Beveridge, D. L.; Dobosh, P. A. *J. Chem. Phys.* **1967**, *47*, 2026.
(16) Hinchliffe, A. *Modelling Molecular Structures*; 2nd ed.; John Wiley and Sons, 2000.
(17) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; John Wiley & Sons, Inc., 2001.
(18) Abdulnur, S. F.; Laki, K. *Biophys. J.* **1979**, *28*, 503.

(19) Zerner, M. C.; Ridley, J. *Theor. Chem. Acc.* **1973**, *32*, 111.
(20) Coffey, P.; Jug, K. *J. Am. Chem. Soc.* **1973**, *95*, 7575.
(21) Jug, K. *Theor. Chem. Acc.* **1976**, *42*, 303.
(22) Nanda, D. N.; Jug, K. *Theor. Chem. Acc.* **1980**, *57*, 95.
(23) Pilar, F. L. *Elementary Quantum Chemistry*; 2nd ed.; Dower Publications, New York, 1990.
(24) Bingham, R. C.; Dewar, M. J. S.; Lo, D. H. *J. Am. Chem. Soc.* **1975**, *97*, 1285.
(25) Baird, N. C.; Dewar, M. J. S. *J. Chem. Phys.* **1969**, *50*, 1262.
(26) Dewar, M. J. S.; Haselbach, E. *J. Am. Chem. Soc.* **1970**, *92*, 590.
(27) Fletcher, R.; Powell, M. J. D. *Computer J.* **1963**, *6*, 163.
(28) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(29) Thiel, W. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; Wiley, Chichester: 1998; Vol. 3, p 1599.
(30) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(31) Lopez, X.; York, D. M. *Theor. Chem. Acc.* **2003**, *109*, 149.
(32) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, *9*, 408.
(33) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(34) Kayi, H.; Clark, T. *J. Mol. Model.* **2011**, *17*, 2585.
(35) Jensen, F. *Introduction to Computational Chemistry*; 2nd ed.; John Wiley and Sons, Ltd., 2007.
(36) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
(37) Dewar, M. J. S.; Healy, E. F.; Holder, A. J.; Yuan, Y.-C. *J. Comput. Chem.* **1990**, *11*, 541.
(38) Stewart, J. J. P. *J. Comput. Chem.* **1990**, *11*, 543.
(39) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391.
(40) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, *93*, 315.
(41) Dewar, M. J. S.; Thiel, W. *Theor. Chim. Acta (Berl.)* **1977**, *46*, 89.
(42) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616.
(43) Dewar, M. J. S.; Jie, C.; Yu, G. *Tetrahedron* **1993**, *49*, 5003.
(44) Cramer, C. J. *Essentials of Computational Chemistry*; 2nd ed.; John Wiley and Sons, Ltd., 2004.
(45) Holder, A. J. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; Wiley, Chichester: 1998; Vol. 3, p 2542.
(46) AMPAC; 9.0: Semichem, 12456 W, 62nd Terrace, Shawnee, KS 66216, 2008.
(47) Hehre, W. J. *Practical Strategies for Electronic Structure Calculations*; Wavefunction, 1995.
(48) SPARTAN; 8.0: Wavefunction Inc., 18401 Von Karman Avenue, Irvine, CA 92715, 2008.
(49) Voityuk, A. A.; Rösch, N. *J. Phys. Chem. A* **2000**, *104*, 4089.
(50) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1994**, *116*, 3892.
(51) Cativiela, C.; Dillet, V.; García, J. I.; Mayoral, J. A.; Ruiz-López, M. F.; Salvatella, L. *J. Mol. Struct. (Theochem)* **1995**, *331*, 37.

(52) Khalil, M.; Woods, R. J.; Weaver, D. F.; Smith, V. H. *J. Comput. Chem.* **1991**, *12*, 584.
(53) Csonka, G. I. *J. Comput. Chem.* **1993**, *14*, 895.
(54) Csonka, G. I.; Angyan, J. G. *J. Mol. Struct. (Theochem)* **1997**, *393*, 31.
(55) Bernal-Uruchurtu, M. I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-López, M. F. *J. Comput. Chem.* **2000**, *21*, 572.
(56) Bernal-Uruchurtu, M. I.; Ruiz-López, M. F. *Chem. Phys. Lett.* **2000**, *330*, 118.
(57) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601.
(58) Tubert-Brohman, I.; Guimaraes, C. R. W.; Repasky, M. P.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 138.
(59) Tubert-Brohman, I.; Guimaraes, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817.
(60) Winget, P.; Clark, T. *J. Mol. Model.* **2005**, *11*, 439.
(61) Kayi, H.; Clark, T. *J. Mol. Model.* **2007**, *13*, 965.
(62) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 295.
(63) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 1253.
(64) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 29.
(65) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 1109.
(66) McNamara, J. P.; Muslim, A.-M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, *394*, 429.
(67) Barrows, S. E.; Dulles, F. J.; Cramer, C. J.; French, A. D.; Truhlar, D. G. *Carbohydr. Res.* **1995**, *276*, 219.
(68) Appell, M.; Strati, G.; Willett, J. L.; Momany, F. A. *Carbohydr. Res.* **2004**, *339*, 537.
(69) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231.
(70) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
(71) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(72) Stewart, J. J. P.; Stewart Computational Chemistry: MOPAC2009, Colorado Springs, CO, USA, http://openmopac.net, 2008.
(73) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.
(74) Kolb, M.; Thiel, W. *J. Comput. Chem.* **1993**, *14*, 775.
(75) Kolb, M. PhD Thesis, Universität Wuppertal, 1991.
(76) Weber, W. PhD Thesis, Universität Zurich, 1996.
(77) Scholten, M. PhD Thesis, Universität Düsseldorf, 2003.
(78) Silva-Junior, M. R.; Thiel, W. *J. Chem. Theory Comput.* **2010**, *6*, 1546.
(79) Mane, J. Y.; Klobukowski, M. *Chem. Phys. Lett.* **2010**, *500*, 140.
(80) Stewart, J. J. P. *J Mol Model* **2013**, *19*, 1.
(81) Stewart, J. J. P.; Stewart Computational Chemistry: MOPAC2012, Colorado Springs, CO, USA, http://openmopac.net, 2012.
(82) Foster, M. E.; Sohlberg, K. *J. Chem. Theory Comput.* **2010**, *6*, 2153.
(83) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362.
(84) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.
(85) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28*, 555.
(86) Korth, M.; Pitonak, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.
(87) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808.
(88) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141.

(89) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
(90) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.
(91) Goringe, C. M.; Bowler, D. R.; Hernandez, E. *Rep. Prog. Phys.* **1997**, *60*, 1447.
(92) Bersuker, I. B. In *Computational Chemistry: Reviews of Current Trends*; Leszczynski, J., Ed.; World Scientific Publishing Co. Pte. Ltd.: 2001; Vol. 6, p 69.
(93) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
(94) Senn, H. M.; Thiel, W. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.
(95) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700.
(96) Matsubara, T.; Maseras, F.; Koga, N.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 2573.
(97) Genest, A.; Woiterski, A.; Krüger, S.; Shor, A. M.; Rösch, N. *J. Chem. Theory Comput.* **2006**, *2*, 47.
(98) Eichler, U.; Kölmel, C. M.; Sauer, J. *J. Comput. Chem.* **1997**, *18*, 463.
(99) To, J.; Sherwood, P.; Sokol, A. A.; Bush, I. J.; Catlow, C. R. A.; van Dam, H. J. J.; French, S. A.; Guest, M. F. *J. Mater. Chem.* **2006**, *16*, 1919.
(100) Cembran, A.; Gao, J. *Mol. Phys.* **2006**, *104*, 943.
(101) Tubert-Brohman, I.; Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 16904.
(102) Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128*, 6141.
(103) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.
(104) Senn, H. M.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173.
(105) Grotendorst, J.; Attig, N.; Blüegel, S.; Marx, D. *Multiscale Simulation Methods in Molecular Sciences* Julich Supercomputing Centre, 2009; Vol. 42.
(106) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283.
(107) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161.
(108) Yu, H.; van Gunsteren, W. F. *Comput. Phys. Commun.* **2005**, *172*, 69.
(109) Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory Comput.* **2007**, *3*, 2034.
(110) Gao, J. *J. Comput. Chem.* **1996**, *18*, 1061.
(111) Thompson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492.
(112) Lin, Y.-L.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1484.
(113) Bryce, R. A.; Vincent, M. A.; Malcolm, N. O. J.; Hillier, I. H.; Burton, N. A. *J. Chem. Phys.* **1998**, *109*, 3077.
(114) Jensen, L.; van Duijnen, P. T.; Snijders, J. G. *J. Chem. Phys.* **2003**, *118*, 514.
(115) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 1499.
(116) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2008**, *10*, 297.
(117) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189.
(118) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, *100*, 10580.
(119) Field, M. J. *J. Comput. Chem.* **2002**, *23*, 48.
(120) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.
(121) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170.
(122) Lyne, P. D.; Hodoscek, M.; Karplus, M. *J. Phys. Chem. A* **1999**, *103*, 3462.
(123) Amara, P.; Field, M. J. *Theor. Chem. Acc.* **2003**, *109*, 43.
(124) Zhang, Y. *Theor. Chem. Acc.* **2006**, *116*, 43.
(125) Ferré, N.; Olivucci, M. *J. Mol. Struct. (Theochem)* **2003**, *632*, 71.

(126) Théry, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269.
(127) Monard, G.; Loos, M.; Théry, V.; Baka, K.; Rivail, J.-L. *Int. J. Quantum Chem.* **1996**, *58*, 153.
(128) Assfeld, X.; Rivail, J.-L. *Chem. Phys. Lett.* **1996**, *263*, 100.
(129) Gorb, L. G.; Rivail, J.-L.; Théry, V.; Rinaldi, D. *Int. J. Quantum Chem.* **1996**, *30*, 1525.
(130) Murphy, R. B.; M., P. D.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442.
(131) Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **1999**, *20*, 1468.
(132) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *Chem. Phys. Lett.* **2000**, *321*, 113.
(133) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 5454.
(134) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 632.
(135) Amara, P.; Field, M. J.; Alhambra, C.; Gao, J. *Theor. Chem. Acc.* **2000**, *104*, 336.
(136) Steinbrecher, T.; Elstner, M. In *Biomolecular Simulations: Methods and Protocols (Methods in Molecular Biology)*; Monticelli, L., Salonen, E., Eds.; Springer Science and Business Media: New York, 2013; Vol. 924, p 91.
(137) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, 1989.
(138) Leach, A. R. *Molecular Modelling: Principles and applications*; 2nd ed.; Prentice Hall, 2001.
(139) Haile, J. M. *Molecular Dynamics Simulation: Elementary Methods*; Wiley, 1997.

3. Quantum Mechanics

In Chapter 2 a historical background to various quantum mechanical methods was provided. In this chapter the theoretical background surrounding these methods is considered. Primary focus is placed on the theory surrounding semi-empirical based methods and the corrections which can be utilized to increase the accuracy of such methods.

3.1 General approximations

The most important goal of many approaches made in quantum chemistry is to approximately solve the time-independent, non-relativistic Schrödinger equation,

$$H\Psi = E\Psi , \tag{3.1}$$

where H is the Hamiltonian for a system (a molecule), E is the energy of the system and Ψ is the wavefunction containing all information that can possibly be known about the quantum chemical system. Mathematically, H is considered to be an operator, E is an eigenvalue that is a scalar value and Ψ is an eigenfunction.

The theory for which we shall provide more detail is one which forms the basis of all quantum mechanical (QM) methods, known as the *Hartree–Fock* (HF) theory (or *Self-Consistent-Field* approximation). Within this theory the total Hamiltonian operator can be written as the sum of kinetic and potential energies of the nuclei and electrons,

$$H_{tot} = T_n + T_e + V_{ne} + V_{ee} + V_{nn} , \tag{3.2}$$

where T_n and T_e are the kinetic energy terms of the nuclei and electrons, respectively. V_{ne}, V_{ee}, and V_{nn} are potential energy terms of the nucleus–electron attraction, electron–electron repulsion and nucleus–nucleus repulsion, respectively.

The nuclear kinetic energy term is a sum of differential operators,

$$T_n = \sum_{A=1}^{N_{nuclei}} -\frac{1}{2M_A}\nabla_A^2 , \tag{3.3}$$

where M_A is the mass of nucleus A and ∇^2 is the Laplacian operator, which is specific to each particle and if one works in Cartesian coordinates is defined as,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{dy^2} + \frac{\partial^2}{\partial z^2}. \tag{3.4}$$

The remaining terms of the Hamiltonian operator can be expanded as follows,

$$T_e = -\sum_{i=1}^{N_{elec}} \frac{1}{2} \nabla_i^2 , \tag{3.5}$$

$$V_{ne} = -\sum_{A=1}^{N_{nuclei}} \sum_{i=1}^{N_{elec}} \frac{Z_A}{R_{Ai}}, \tag{3.6}$$

$$V_{ee} = \sum_{i=1}^{N_{elec}} \sum_{j>i}^{N_{elec}} \frac{1}{r_{ij}}, \tag{3.7}$$

$$V_{nn} = \sum_{A=1}^{N_{nuclei}} \sum_{B>A}^{N_{nuclei}} \frac{Z_A Z_B}{R_{AB}}, \tag{3.8}$$

where the indices i and j indicate the electrons and A and B indicate the nuclei. R_{Ai} is the distance between atom A and electron i and j, r_{ij} is the distance between electrons i and j, R_{AB} is the distance between atoms A and B, respectively. Note that the zero point of the energy corresponds to the particles being at rest ($T_e = 0$) and infinitely removed from one another ($V_{ne} = V_{ee} = V_{nn} = 0$).

Here some simplification to the Schrödinger equation is applied. Since the mass of the nucleus is much heavier than that of the electron, nuclei move slower than the electrons. Thus, electrons can be considered to be moving in a field of fixed nuclei.[1] The nuclear–nuclear repulsion does not depend on electron coordinates and is a constant for a given nuclear geometry and since the nuclei are considered to be stationary in space, their kinetic energy becomes zero. This approximation is known as the Born–Oppenheimer approximation.[2] By applying this

approximation, the Hamiltonian operator can be separated into nuclear and electronic Hamiltonian parts,

$$H_{tot} = H_{nuc} + H_{elect}, \tag{3.9}$$

with H_{elect} being expressed as follows,

$$H_{elect} = \sum_{i=1}^{N_{elec}} h(i) + \sum_{i=1}^{N_{elec}} \sum_{j>i}^{N_{elec}} g(i,j), \tag{3.10}$$

where *h(i)* is the one-electron operator describing the motion of electron *i* in the field of all the nuclei and *g(i, j)* is a two-electron operator giving the electron-electron repulsion,

$$h(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{N_{nuclei}} \frac{Z_A}{R_{Ai}}, \tag{3.11}$$

$$g(i,j) = \frac{1}{r_{ij}}. \tag{3.12}$$

The approximation mentioned above (Born–Oppenheimer) also allows us to write the total energy as the sum of electronic energy and constant nuclear repulsion,

$$E_{tot} = E_{nuc} + E_{elect} = V_{nn} + E_{elect}, \tag{3.13}$$

where V_{nn} is given by eq. 3.8.

From here we need to calculate the electronic energy, E_{elect}, using the electronic wavefunction, Ψ(r; R) and electronic Hamiltonian, H_{elect}. This results in the following equation,

$$H_{elect}\,\Psi(r;R) = E_{elect}\,\Psi(r;R), \tag{3.14}$$

with r and R denoting the electronic and nuclear degrees of freedom, respectively.

At this point it is important to note that the Schrödinger equation is exactly solvable only for one electron systems, such as the hydrogen atom. However, for two (or in general many electron systems) assuming that electrons do not interact with each other gives a Hamiltonian that is separable and the total electronic wavefunction, $\Psi(r_1, r_2)$, describing the motions of the

two electrons would just be the product of two hydrogen atom wavefunctions (orbitals), $\Psi_H(r_1)\Psi_H(r_2)$.[3]

Assuming that the electrons do not interact is a considerably bold approximation, to say the least. Nevertheless, we have to start somewhere and it's plausible to start with a wavefunction of the general form,

$$\Psi_{HP}(r_1, r_2, ..., r_N) = \phi_1(r_1)\phi_2(r_2)...\phi_N(r_N), \tag{3.15}$$

which is known as the *Hartree Product*.

While this functional form is fairly convenient, it has at least one major drawback in that it fails to satisfy the antisymmetry principle (leading to the Pauli exclusion principle), which states that a wavefunction describing fermions (particles having a spin of ½) should be antisymmetric with respect to the interchange of any set of space-spin coordinates.[4] Space-spin coordinates mean that fermions have not only three spatial degrees of freedom, but also an intrinsic spin coordinate, called α of β. A generic (with ω either α or β) set of space-spin coordinates is described by,

$$x = \{r, \omega\}, \tag{3.16}$$

where r is the vector position of a particular electron and ω is the spin coordinate. Thus, the electronic wavefunction becomes a function of 4N variables consisting of the three coordinates and the spin for each electron.

At this point we will change the notation for orbitals from $\phi(r)$, a spatial orbital, to $\chi(x)$, a spin orbital. Therefore, the *Hartree Product* now becomes,

$$\Psi_{HP}(x_1, x_2, ...x_n) = \chi_1(x_1)\chi_2(x_2)...\chi_N(x_N). \tag{3.17}$$

This wavefunction does not satisfy the Pauli exclusion principle, as postulated by Fock in 1930.[5] John Slater[6,7] later went on to express the wavefunction, according to Fock's suggestion, as a determinant, written as follows,

$$\Psi_{Slater} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \chi_3(x_1) & \dots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \chi_3(x_2) & \dots & \chi_N(x_2) \\ \chi_1(x_3) & \chi_2(x_3) & \chi_3(x_3) & \dots & \chi_N(x_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_N) & \chi_2(x_N) & \chi_3(x_N) & \dots & \chi_N(x_N) \end{vmatrix}, \tag{3.18}$$

where χ_i indicates the spin orbitals, N is the total number of electrons and $\frac{1}{\sqrt{N!}}$ is a normalization factor. Ψ_{Slater} is known as a Slater determinant that indicates a Hartree–Fock or SCF wavefunction.

Since we can always construct a determinant (within a sign) if we just know the list of occupied orbitals $\{\chi_i(x)\chi_j(x)\dots\chi_k(x)\}$, we can write it in a ket symbol as $|\chi_i\chi_j\dots\chi_k\rangle$ or even more simply as $|ij\dots k\rangle$.[3] Note that the normalization factor has not been explicitly included, but is implied.

Now we move onto determining the HF energy expression. In order to do this we first consider the HF wavefunction as having the form of a Slater determinant, which will then produce an expectation energy that is given by the usual quantum mechanical expression (assuming the wavefunction is normalized),

$$E_{elect} = \langle \Psi | H_{elect} | \Psi \rangle. \tag{3.20}$$

For symmetric energy expression we employ the variational principle, which states that the best wavefunction is the one with the lowest energy. Hence, better approximate wavefunctions can be obtained by varying their parameters until the energy is minimized within the given functional space. Therefore, the correct molecular orbitals are those which minimize the electronic energy (E_{elect}). The molecular orbitals can be obtained numerically using integration over a grid, or (as is more common) be represented as a linear combination of atomic orbitals (LCAO).[8-10]

The next step involves substituting the electronic Hamiltonian with the one- and two-electron operators provided in eqs. 3.11 and 3.12, respectively. After some manipulation and simplification, allocated in various textbooks,[11-14] we obtain the HF energy expression,

$$E_{HF} = \sum_{i}^{N_{elec}} \langle i \mid h \mid i \rangle + \frac{1}{2} \sum_{ij}^{N_{elec}} \langle ij \mid ij \rangle - \langle ij \mid ji \rangle, \tag{3.21}$$

where the one-electron integral is,

$$\langle i \mid h \mid j \rangle = \int dx_1 \chi_i^*(x_1) h(r_1) \chi_j(x_1), \tag{3.22}$$

and the two-electron integral is,

$$\langle ij \mid kl \rangle = \int dx_1 dx_2 \chi_i^*(x_1) \chi_j(x_1) \frac{1}{r_{12}} \chi_k^*(x_2) \chi_l(x_2). \tag{3.23}$$

Since the above mentioned energy expression is symmetric, the variational theorem holds, and so we know that the Slater determinant with the lowest energy is as close as we can get to the true wavefunction for the assumed functional form of a single Slater determinant. The HF method determines the set of spin orbitals which minimize the energy and give us this "best single determinant".[3] With this being said we need to now minimize the HF energy expression with respect to changes in the orbitals ($\chi_i \rightarrow \chi_i + \delta\chi_i$). Up to this point we have assumed that the orbitals (χ) are orthonormal, which is something we now need to ensure is still the case after application of the variational principle. This can be accomplished by Lagrange's method of undetermined multipliers, where a function L is employed,

$$L[\{\chi_i\}] = E_{HF}[\{\chi_i\}] - \sum_{ij} \varepsilon_{ij} \left(\langle i \mid j \rangle - \delta_{ij} \right), \tag{3.24}$$

where ε_{ij} are the undetermined Lagrange multipliers and $\langle i \mid j \rangle$ is the overlap between spin orbitals i and j, which is expressed as,

$$\langle i \mid j \rangle = \int \chi_i^*(x) \chi_j(x) dx. \tag{3.25}$$

Setting the first variation δL=0, and working through some algebra, we eventually arrive at the HF equations defining the orbitals,

$$\begin{aligned} & h(x_1)\chi_i(x_1) + \sum_{j \neq i} \left[\int dx_2 \mid \chi_j(x_2) \mid^2 r_{12}^{-1} \right] \chi_i(x_1) \\ & - \sum_{j \neq i} \left[\int dx_2 \chi_j^*(x_2) \chi_i(x_2) r_{12}^{-1} \right] \chi_j(x_1) = \varepsilon_i \chi_i(x_1) \end{aligned}, \tag{3.26}$$

where ε_i is the energy eigenvalue associated with orbital χ_i.

The HF equations can be solved numerically (exact HF),[12] or they can be solved in the spaced spanned by a set of basis functions (HF-Roothaan equations).[15,16] In either case, note that the solutions depend on the orbitals. Hence, we need to guess some initial orbitals and then refine the guesses iteratively. It is for this reason that HF is called a self-consistent-field (SCF) approach.

At this point it is a good idea for us to redefine the terms provided in eq. 3.26 above. The first term given in square brackets provides the *Coulomb* interaction of an electron in spin orbital χ_i with the average charge distribution of the other electrons. This is called the *Coulomb* term, and a convenient way to define a *Coulomb* operator is,

$$J_j(x_1) = \int dx_2 \mid \chi_j(x_2) \mid^2 r_{12}^{-1}, \tag{3.27}$$

which gives the average local potential at point x_i due to the charge distribution from the electron in orbital χ_j.

The second square bracketed term in eq. 3.26 is a bit more complicated to explain and does not have a simple classical analog. It arises from the antisymmetry requirement of the wavefunction. It looks much like the Coulomb term, except that it switches or exchanges spin orbitals χ_i and χ_j. Hence, it is called the *exchange* term and the *exchange* operator can be defined in terms of its action on an arbitrary spin orbital χ_i,

$$K_j(x_1)\chi_i(x_1) = \left[\int dx_2 \chi_j^*(x_2) r_{12}^{-1} \chi_i(x_2) \right] \chi_j(x_1). \tag{3.28}$$

Based on the *Coulomb* and *exchange* operators the HF equations become considerably more compact,

$$\left[h(x_1) + \sum_{j \neq i} J_j(x_1) - \sum_{j \neq i} K_j(x_1) \right] \chi_i(x_1) = \varepsilon_i \chi_i(x_1). \tag{3.29}$$

If we now realize that,

$$\left[J_i(x_1) - K_i(x_1) \right] \chi_i(x_1) = 0, \tag{3.30}$$

then it becomes clear that we can remove the restrictions $j \neq i$ in the summations, and we can introduce a new operator known as the *Fock operator*,

$$f(x_1) = h(x_1) + \sum_j J_j(x_1) - K_j(x_1). \tag{3.31}$$

And now we have an even more simplified form for the HF equations,[3]

$$f(x_1)\chi_i(x_1) = \varepsilon_i \chi_i(x_1) \tag{3.32}$$

Introducing a basis set transforms the HF equations into the Roothaan equations.[15,16] Denoting the atomic orbital basis function as φ, we obtain an expression,[12]

$$\chi_i = \sum_{\mu=1}^{N_{AOs}} c_{\mu i}\varphi_\mu , \tag{3.33}$$

where N_{AOs} is the number of atomic orbitals in the system and $c_{\mu i}$ is the coefficient of atomic orbital φ_μ. This then leads to,

$$f(x_1)\sum_\mu c_{\mu i}\varphi_\mu(x_1) = \varepsilon_i \sum_\mu c_{\mu i}\varphi_\mu(x_1). \tag{3.34}$$

Multiplying the left hand side of the above equation by $\varphi_\nu^*(x_1)$ and integrating yields a matrix equation,

$$\sum_\mu c_{\mu i}\int dx_1 \varphi_\mu(x_1) f(x_1)\varphi_\nu^*(x_1) = \varepsilon_i \sum_\mu c_{\mu i}\int dx_1 \varphi_\mu(x_1)\varphi_\nu^*(x_1). \tag{3.35}$$

The equation above can be simplified by introducing a general matrix element notation,[11,12]

$$S_{\mu\nu} = \int dx_1 \varphi_\mu(x_1)\varphi_\nu^*(x_1), \tag{3.36}$$

$$F_{\mu\nu} = \int dx_1 \varphi_\mu(x_1) f(x_1)\varphi_\nu^*(x_1). \tag{3.37}$$

Now the HF-Roothaan equations can be written in matrix form as,

$$\sum_\mu F_{\mu\nu} c_{\mu i} = \varepsilon_i \sum_\mu S_{\mu\nu} c_{\mu i} , \tag{3.38}$$

or in an even more simplified notation as matrices,

$$FC = SC\varepsilon , \tag{3.39}$$

where ε is a diagonal matrix of the orbital energies ε_i.

This is like an eigenvalue equation except for the overlap matrix (**S**). Transforming the basis to one that is orthogonal would make **S** vanish. It is then just a matter of solving an eigenvalue equation, and since **F** depends on its own solution (through the orbitals), the process must be done iteratively. This is why the solution of the HF-Roothaan equations are often called the SCF procedure.[3]

At this point it is important that one realizes that the HF model is a kind of branching point, where either additional approximations can be invoked, leading to semi-empirical (SE) methods, or it can be improved by adding additional determinants, thereby generating models that can be made to converge towards the exact solution of the electronic Schrödinger equation.[12] A schematic diagram depicting the branching of the HF method is provided in Figure 3.1 below.

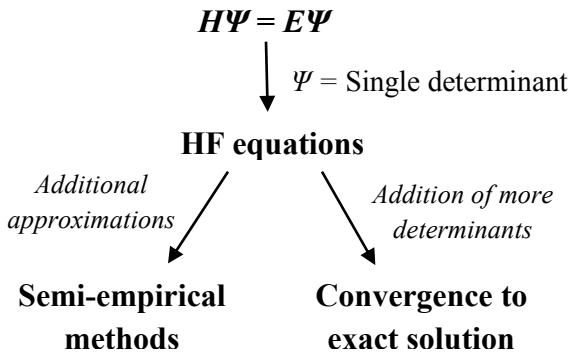


Figure 3.1: HF model as a starting point for more approximate or more accurate treatments.

In the sections that follow we shall look at the theory surrounding various SE based methods as well as one of the various methods in which additional determinants are added to the HF model, known as Density Functional Theory (DFT).

3.2 Semi-empirical Methods

Semi-empirical (SE) methods, in general, can be distinguished from each other by:[1]

i) The electrons being treated (e.g., models based on π-electron or all valence-electron treatment).

ii) The system to which they can be applied (e.g., linear, branched or planar conjugate systems).

iii) The differential overlap being neglected (e.g., CNDO, INDO or NDDO approximations).

iv) The treatment of the core-core repulsion term (e.g., use of atom specific or atom-pair specific parameters in the core repulsion function).

v) The parameterization strategy (e.g., chemical intuition or fully automated).

vi) The values of parameters (e.g., methods using the same theoretical foundation and same approximations such as AM1 and RM1).

As mentioned in Chapter 2 (Section 2.2) there are various SE methods in existence, but in the sections that follow we shall focus particularly on the theoretical frame of NDDO type methods (also known as MNDO-like methods) as these methods are the most widely used and most popular to date.

3.2.1 MNDO

The work of Dewar and Thiel in 1977[17] yielded the Modified Neglect of Differential Overlap (MNDO) method which evaluates one-center two-electron integrals based on spectroscopic data for isolated atoms, and evaluates other two-electron integrals using the idea of multipole-multipole interactions from classical electrostatics. With this method various integrals are based on numerical parameters that are adjusted to fit experimental data, rather than being determined analytically.

Since the work conducted in this thesis pertains to closed-shell systems we shall describe the MNDO formalism based on this context. For closed-shell systems MNDO uses the frozen core approximation (core electrons are considered as part of the nucleus, so that the electronic energy expression explicitly involves only the valence electrons). In addition, the valence shell molecular orbitals χ_i and the corresponding orbital energies ε_i are obtained from the solution of the secular equations,[1]

$$\chi_i = \sum_{\nu=1} c_{\nu i} \varphi_\nu \ , \tag{3.40}$$

where φ_v represents the atomic orbitals of valence electrons. The coefficients c_{vi} are calculated from the Roothaan-Hall[15,16] equations, that for the NDDO approximation (where overlap $S_{\mu\nu} = \delta_{\mu\nu}$) take on the following form,

$$\sum_{\nu}\left(F_{\mu\nu} - \delta_{\mu\nu}\varepsilon_i\right)c_{\nu i} = 0, \tag{3.41}$$

where ε_i is the Eigenvalue of molecular orbital χ_i, and $\delta_{\mu\nu}$ is the Kronecker-delta (equal to one if $\mu = \nu$ and zero otherwise).

If φ_μ is the same as φ_v then, because of the symmetry of the two-electron integrals, the diagonal element is written as,

$$F^{\alpha}_{\mu\mu} = H_{\mu\mu} + \sum_{\nu}^{A}\left[P^{\alpha+\beta}_{\nu\nu}\left\langle\mu\mu\,|\,\nu\nu\right\rangle - P^{\alpha}_{\nu\nu}\left\langle\mu\nu\,|\,\mu\nu\right\rangle\right] + \sum_{B\neq A}^{B}\sum_{\lambda}^{B}\sum_{\sigma}P^{\alpha+\beta}_{\lambda\sigma}\left\langle\mu\mu\,|\,\lambda\sigma\right\rangle, \tag{3.42}$$

where $H_{\mu\mu}$ represents the energy an electron in atomic orbital φ_μ would have if all electrons were removed from the system and is given as,

$$H_{\mu\mu} = U_{\mu\mu} - \sum_{B\neq A}Z_B\left\langle\mu\mu\,|\,BB\right\rangle, \tag{3.43}$$

where $U_{\mu\mu}$ is the one-electron energy (obtained parameterically), Z_B is the effective charge of atom B, and $\langle\mu\mu\,|\,BB\rangle$ is the core-electron integral.

By equating the core-electron integral to the corresponding two-electron integral one obtains,

$$\left\langle\mu\mu\,|\,BB\right\rangle = \left\langle\mu_A\mu_A\,|\,s_B s_B\right\rangle, \tag{3.44}$$

which then results in $H_{\mu\mu}$ being rewritten as,

$$H_{\mu\mu} = U_{\mu\mu} - \sum_{B\neq A}Z_B\left\langle\mu_A\mu_A\,|\,s_B s_B\right\rangle. \tag{3.45}$$

If φ_μ and φ_v are different, but on the same center, then since a minimal basis is used for NDDO type methods, all integrals of the type $\langle\mu\nu\,|\,\lambda\sigma\rangle$ are zero by the orthogonality of the atomic orbitals unless $\mu = \nu$ and $\lambda = \sigma$ or $\mu = \lambda$ and $\nu = \sigma$.[18] The off-diagonal elements are represented as,

$$F^{\alpha}_{\mu\nu} = H_{\mu\nu} + 2P^{\alpha+\beta}_{\mu\nu}\left\langle\mu\nu\,|\,\mu\nu\right\rangle - P^{\alpha}_{\mu\nu}\left[\left\langle\mu\nu\,|\,\mu\nu\right\rangle + \left\langle\mu\mu\,|\,\nu\nu\right\rangle\right]. \tag{3.46}$$

In the case were integrals $\langle\mu\nu\,|\,\lambda\sigma\rangle$ are not zero, the off-diagonal elements are given as,

$$F_{\mu\nu}^{\alpha} = H_{\mu\nu} + 2P_{\mu\nu}^{\alpha+\beta}\langle\mu\nu\,|\,\mu\nu\rangle - P_{\mu\nu}^{\alpha}\left[\langle\mu\nu\,|\,\mu\nu\rangle + \langle\mu\mu\,|\,\nu\nu\rangle\right] + \sum_{B\neq A}\sum_{\lambda}^{B}\sum_{\sigma}^{B} P_{\lambda\sigma}^{\alpha+\beta}\langle\mu\nu\,|\,\lambda\sigma\rangle, \quad (3.47)$$

where $H_{\mu\nu}$ is the two-center one-electron integral (resonance integral) which is approximated using the overlap integral $S_{\mu\nu}$ given as,

$$S_{\mu\nu} = \langle\varphi_{\mu}\varphi_{\nu}\rangle. \quad (3.48)$$

The resulting resonance integral is then written as,

$$H_{\mu\nu} = \frac{1}{2} S_{\mu\nu}\left(\beta_{\mu} + \beta_{\nu}\right), \quad (3.49)$$

where β_{μ} is an adjustable parameter that is characteristic of the φ_{μ} atomic orbital at atom A, and β_{ν} is an adjustable parameter of the φ_{ν} atomic orbital at atom B.

The resonance integral contributes mainly to the bonding energy of a molecule. In addition, as far as first-row elements are concerned, there are at most only two different β parameters (β_s and β_p) to be optimized since these atoms contain only s- and p-orbitals. However, for nitrogen and oxygen these two parameters are set equal ($\beta_s = \beta_p$) and not optimized separately.[1]

Together with the equations given above, one must also consider the Slater-orbital exponents (ζ_s and ζ_p), which for elements that possess s- and p-orbitals are set equal ($\zeta_s = \zeta_p$). In addition one-center two-electron repulsion integrals (G_{ss}, G_{pp}, G_{p2}, G_{sp}, H_{sp}), based on the multipole expansion, are derived from experimental data on isolated atoms. Most of this data was taken from work by Oleari et al.,[19] but some of the data was obtained by optimization to fit molecular properties. For all atoms, except hydrogen and helium, there are a maximum of five one-center two-electron integrals, given as,

$$G_{ss} = \langle ss\,|\,ss\rangle, \quad (3.50)$$

$$G_{sp} = \langle ss\,|\,pp\rangle, \quad (3.51)$$

$$H_{sp} = \langle sp\,|\,sp\rangle, \quad (3.52)$$

$$G_{pp} = \langle pp\,|\,pp\rangle, \quad (3.53)$$

$$G_{p2} = \langle pp\,|\,p'p'\rangle, \quad (3.54)$$

where p and p' are two different p-type atomic orbitals.

Using the definitions above, the one-center two-electron contributions to the Fock matrix become,[18]

$$F_{ss}^{\alpha} : P_{ss}^{\beta} G_{ss} + \left(P_{px}^{\alpha+\beta} + P_{py}^{\alpha+\beta} + P_{pz}^{\alpha+\beta}\right) G_{sp} - \left(P_{px}^{\alpha} + P_{py}^{\alpha} + P_{pz}^{\alpha}\right) H_{sp}, \tag{3.55}$$

$$F_{sp}^{\alpha} : 2P_{sp}^{\alpha+\beta} H_{sp} - P_{sp}^{\alpha}\left(H_{sp} + G_{sp}\right), \tag{3.56}$$

$$F_{pp}^{\alpha} : P_{ss}^{\alpha+\beta} G_{sp} - P_{ss}^{\alpha} H_{sp} + P_{pp}^{\beta} G_{pp} + \left(P_{p'}^{\alpha+\beta} + P_{p''}^{\alpha+\beta}\right) G_{p2} - \frac{1}{2}\left(P_{p'}^{\alpha} + P_{p''}^{\alpha}\right)\left(G_{pp} - G_{p2}\right), \tag{3.57}$$

$$F_{pp'}^{\alpha} : P_{pp'}^{\alpha+\beta}\left(G_{pp} - G_{p2}\right) - \frac{1}{2} P_{pp'}^{\alpha}\left(G_{pp} + G_{p2}\right). \tag{3.58}$$

For purposes of this thesis we shall *not* look at the 22 two-center two-electron integrals which exist for each pair of heavy (non-hydrogen) atoms, as this can be found elsewhere.[18] What we shall consider, however, is the determination of the total energy (E_{TOT}) of a molecule, which is represented by the sum of its electronic energy (E_{el}), and the repulsion energy (E_{AB}^{CORE}) between the cores of atom A and B. The total energy is written as,

$$E_{TOT} = E_{el} + \sum \sum_{A<B} E_{AB}^{CORE} , \tag{3.59}$$

where the electronic energy is given as,

$$E_{el} = \frac{1}{2} \sum_{\mu} \sum_{\nu} P_{\mu\nu}\left(H_{\mu\nu} + F_{\mu\nu}\right), \tag{3.60}$$

where $P_{\mu\nu}$ is the density matrix, $H_{\mu\nu}$ is the one-electron part of the core Hamiltonian and $F_{\mu\nu}$ is the Fock matrix.

The repulsion energy term (E_{AB}^{CORE}), more commonly known as the core-core repulsion energy term, of the MNDO method has the following form,

$$E_{AB}^{CORE} = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle + f(R_{AB}), \tag{3.61}$$

where $f(R_{AB})$ is expanded as,

$$f(R_{AB}) = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle \left[e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right]. \tag{3.62}$$

In cases where either O-H or N-H interactions are experienced, it was found advantages to make use of a slight modification to the term given above,

$$f(R_{AH}) = Z_A Z_H \left\langle s_A s_A \mid s_H s_H \right\rangle \left[R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}} \right]. \tag{3.63}$$

By combining the equations above the MNDO core-core repulsion term can be written as,

$$E_{AB}^{MNDO} = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle \left[1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right], \tag{3.64}$$

and if A = N or O and B = H,

$$E_{AH}^{MNDO} = Z_A Z_H \left\langle s_A s_A \mid s_H s_H \right\rangle \left[1 + R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}} \right], \tag{3.65}$$

where Z is the effective charge of the element, $\left\langle s_A s_A \mid s_X s_X \right\rangle$ (with X = B or H) is the two-center integral of type $\left\langle ss \mid ss \right\rangle$, α is an adjustable element specific parameter, and R_{AY} is the distance between atoms A and Y (with Y = B or H).

Aside from the approximations and parametric functions provided above, MNDO was the first method that could represent lone-pair to lone-pair interactions that were ignored by its predecessors.[1] This made MNDO very popular; however, deficiencies in the method became more apparent as time progressed. Two major problems existed with the method:

i) It was unable to model systems possessing hydrogen bonds accurately.
ii) It was unable to model hypervalent compounds accurately, predicting incorrect energies, geometries and point groups.

Within the sections that follow we shall see how the above mentioned problems were tackled by the MNDO successor methods.

3.2.2 AM1

AM1 is an extension of, a modification to and a re-parameterization of the MNDO method.[20] In essence AM1 differs from MNDO in the following ways:

i) The modification of the core-core repulsion function.

ii) The parameterization of the overlap terms (β_s and β_p), and Slater-type orbital exponents (ζ_s and ζ_p) on the same atom independently, instead of setting them equal as in MNDO.

MNDO has a very strong tendency to overestimate repulsions between atoms when they are within hydrogen bond distances. To overcome this hydrogen bond problem, the net electrostatic repulsion term of MNDO, $f(R_{AH})$ given in eq. 3.63, was modified in MNDO/H[21] to produce,

$$f(R_{AH}) = Z_A Z_H \left\langle s_A s_A \mid s_H s_H \right\rangle \left[e^{-\alpha R_{AH}^2} \right], \tag{3.66}$$

where α is 2.0 Å^{-2} for all A-H pairs.

A slightly different approach was adopted for AM1, where Gaussian functions were added to provide a weak attractive force.[20] The core-core repulsion function of AM1 is then given as,

$$E_{AB}^{AM1} = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle \left[1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}\right] + \frac{Z_A Z_B}{R_{AB}} \left[F(A) + F(B)\right], \tag{3.67}$$

where Gaussian functions $F(A)$ and $F(B)$ are written as,

$$F(A) = \sum_i a_{iA} e^{-b_{iA}(R_{AB} - c_{iA})^2}, \tag{3.68}$$

$$F(B) = \sum_i a_{iB} e^{-b_{iB}(R_{AB} - c_{iB})^2}, \tag{3.69}$$

which results in the final AM1 core-core repulsion function being given as,

$$E_{AB}^{AM1} = E_{AB}^{MNDO} + \frac{Z_A Z_B}{R_{AB}} \left[\sum_i a_{iA} e^{-b_{iA}(R_{AB} - c_{iA})^2} + \sum_i a_{iB} e^{-b_{iB}(R_{AB} - c_{iB})^2} \right]. \tag{3.70}$$

In the equation above the extra terms (a, b, and c) define adjustable spherical Gaussian function parameters. The remaining parameters have the same meaning as in the previous section. Each atom has up to four Gaussian parameters, i.e. $a_1 \ldots a_4$, $b_1 \ldots b_4$, and $c_1 \ldots c_4$. Carbon has four terms

in its Gaussian expansion, whereas hydrogen and nitrogen have three and oxygen has only two terms. The number of Gaussian parameters chosen entails that for carbon, hydrogen and nitrogen both attractive and repulsive Gaussians were used, whereas for oxygen only repulsive ones were considered.

Addition of Gaussian functions into the core-core repulsion term significantly increased the number of parameters to be optimized and made the parameterization process more difficult. With the original MNDO hydrogen and carbon parameters were optimized first followed by the optimization of other elements which were added on one at a time. With AM1 all parameters for H, C, N, and O were optimized at once in a single parameterization procedure. Optimization of these parameters was done manually using chemical knowledge and intuition. The size of reference parameterization data was kept at a minimum by very carefully selecting necessary data to be used as reference.

It is important to note that despite the addition of Gaussian parameters, AM1 (or any NDDO method) does not describe intermolecular interactions very accurately, partially due to the linear interdependence of all two-center interactions[22] and underestimation of molecular polarizabilities.[23] To improve upon this researchers have embarked on applying *p*-type basis functions on the hydrogen atoms, thereby increasing the accuracy of NDDO type methods.[23,24] The theory pertaining to this enhanced polarization is outside the scope of this thesis.

Despite the lack of *p*-orbitals on hydrogen AM1 represented a considerable improvement over MNDO without any increase in the computing time needed. Having been parameterized for many of the main-group elements it is widely used in modeling of organic compounds due to its good performance and robustness.

3.2.3 PM3

In the parameterization of MNDO and AM1, only very few molecules could be used. This was a natural constraint imposed by the software and equipment available at the time these methods were developed.[18] With PM3 a mathematical philosophy for the parameterization procedure was adopted that involved the derivation and implementation of formulae to arrive at a suitable error function with respect to the parameters.[25,26] The error function was given as follows,

$$S = \sum_i (x_i^{calc} - x_i^{ref})^2 , \tag{3.71}$$

where S is defined as the sum of squares of differences between calculated (x_i^{calc}) and reference values (x_i^{ref}). This function is considered *optimized* when for a set of parameters the value of S is a minimum.

Unlike MNDO or AM1, in PM3 the one-center two-electron repulsion parameters (G_{ij}, H_{ij}), given in eqs. 3.50–3.54, are optimized instead of assigning them to atomic spectral values.[1] PM3 does share the core-core repulsion function with AM1, provided in eq. 3.70, however, instead of having up to four Gaussian terms per atom, as in AM1, PM3 only uses two. A total of twelve elements (H, C, N, O, F, Al, Si, P, S, Cl, Br and I) were optimized simultaneously during the original PM3 parameterization.[26] Although PM3 did provide better performance for certain properties when compared to MNDO and AM1, it was not without its deficiencies, which have been provided in Chapter 2 (Section 2.2.7).

3.2.4 MNDO/d

As time progressed computational interests expanded with a number of researchers shifting their attention towards the chemistry surrounding systems which were of a *d*-orbital nature. As such, in 1992, Thiel and Voityuk[27,28] expanded the MNDO formalism to include *d*-orbitals by generalizing the point charge model of MNDO and expanding the two-center two-electron integrals in terms of SE multipole-multipole interactions where all monopoles, dipoles and quadrupoles were included, and all higher order multipoles were neglected.

Aside from changes made to the integrals the authors also looked into the core-electron attractions and core-core repulsions, which in the original MNDO method[17] are given as,

$$V_{\mu\nu,B} = -Z_B \left\langle \mu_A \nu_A \mid s_B s_B \right\rangle , \tag{3.72}$$

$$E_{AB}^{MNDO} = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle \left[1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right], \tag{3.73}$$

where Z_A and Z_B denote the core charges and α refers to element specific parameters.

In eqs. 3.72 and 3.73 the effect of the atomic core is simulated by the valence-shell charge distribution, *ss*, which has no multipole moments higher than the monopole. Although this choice

leads to a realistic balance of the electrostatic interactions[17] in the case of first-row atoms, where the *s*- and *p*-orbitals are generally of comparable size,[27] it is not true for an *spd* basis, where the *s*-, *p*-, and *d*-orbital exponents may be considerably different. As a result MNDO/d has been modified to represent the core by a monopole that is associated with an additive term ρ_{core}. For elements with an *sp* basis $\rho_{core} = \rho_0^{ss}$, where ρ_0^{ss} is given as,

$$\left(\rho_0^{ss}\right)^{-1} = 2G_{ss}. \tag{3.74}$$

This equation is the same as in the original MNDO formalism.[17] For elements with an *spd* basis, however, ρ_{core} is treated as an independent adjustable parameter so that the balance between attractive and repulsive Coulomb interactions is determined by SE parameterization.[27] As such the equations related to the core electrons for MNDO/d are given as,

$$V_{\mu\nu,B} = -Z_B \left\langle \mu_A \nu_A \mid q_{core}^B \right\rangle, \tag{3.75}$$

$$E_{AB}^{MNDO} = Z_A Z_B \left\langle q_{core}^A \mid q_{core}^B \right\rangle \left[1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}\right], \tag{3.76}$$

where the relevant interactions are evaluated according to the point charge model (Section 2 of original MNDO/d paper),[27] e.g.:

$$\left\langle q_{core}^A \mid q_{core}^B \right\rangle = e^2 \left[R_{AB}^2 + \left(\rho_{core}^A + \rho_{core}^B\right)^2\right]^{-1/2}. \tag{3.77}$$

MNDO/d represents a significant improvement over methods such as MNDO, AM1 and PM3.[29]

3.2.5 AM1/d

A short while after the suggestion that further gain over MNDO/d may be expected from an analogous AM1/d parameterization[29] Voityuk and Rösch[30] developed the AM1/d method. The method is an extension of the standard AM1[20] Hamiltonian to an *spd* basis. The established AM1 formalism and the corresponding parameters remain unchanged for all main-group elements therefore, AM1 and AM1/d results are identical for all non-transition metal atoms. As with MNDO/d, the two-center two-electron integrals calculated within an *spd* basis used an extended multipole-multipole interaction scheme and all non-zero one-center two-electron integrals were retained to ensure rotational invariance.[27] For the Molybdenum compounds used during the parameterization of AM1/d it was discovered that inclusion of Gaussian-type functions into the

core-core repulsion term (as was the case for the original AM1, eq. 3.70) did not result in significantly improved results.[30] To overcome this deficiency AM1 was extended by introducing two bond specific parameters (α and δ) into the core-core repulsion term,

$$E_{AB}^{AM1/d} = Z_A Z_B \left\langle s_A s_A \mid s_B s_B \right\rangle \left[1 + \delta_{AB} e^{-\alpha_{AB} R_{AB}}\right]. \tag{3.78}$$

By adopting these new parameters the accuracy of results increased significantly.

Despite the introduction of the bond specific parameters AM1/d did still possess a few deficiencies, which include:

i) Underestimating Mo–O distances by about 0.05 Å.
ii) The angle O=Mo=O of [MoO_2] fragments were calculated smaller than observed.
iii) The double and triple bond lengths of Mo–N were slightly longer when compared to experiment.
iv) Mo–S distances were predicted to be somewhat shorter (by about 0.05 Å) than those found in crystal structures.

Although AM1/d did have the deficiencies mentioned above, it could still be used for computing structural parameters as well as heats of formation, reaction enthalpies, and bond energies of rather large inorganic and organometallic compounds of molybdenum.

3.2.6 AM1*

AM1* was introduced by Winget et al.[31] as an extension to the original AM1 Hamiltonian. The method uses standard MNDO approximations for all integrals involving *s*- and *p*-orbitals and MNDO/d approximations for those including *d*-orbitals.[1] AM1* maintains the parameters of AM1 for elements H, C, N, O, and F. In addition AM1* uses Gaussian functions in the core-core repulsion term as is common for AM1 (eq. 3.70) for the elements H, C, N, O, and F. For all other elements AM1* uses a core-core repulsion term based on the AM1/d Hamiltonian, established by Voityuk and Rösch,[30] as introduction of bond specific parameters was found to be more efficient. However, using these parameters brings the disadvantage of requiring specific parameterization of these terms for every pair of elements. Fortunately, a core-

core repulsion term with these bond specific parameters does not lead to false minima as Gaussian functions can. As such the core-core repulsion term used in AM1* is given as,

$$E_{AB}^{AM1*} = Z_A Z_B \rho_{ss}^0 \left(1 + \delta_{AB} e^{-\alpha_{AB} R_{AB}}\right), \tag{3.79}$$

where α_{AB} and δ_{AB} are bond specific parameters to be optimized, Z_A and Z_B are the effective (valence only) core charges of elements A and B, R_{AB} is the distance between atoms A and B, and $\rho_{ss}^0 = \left\langle q_{core}^A \mid q_{core}^B \right\rangle$, which is the two-center integral defined in the original MNDO/d Hamiltonian[27-29] (eq. 3.77).

In addition, for hydrogen interactions with the elements starting from the second long row of the periodic table the core-core repulsion term of AM1* is represented as,

$$E_{AB}^{AM1*} = Z_A Z_H \rho_{ss}^0 \left(1 + R_{AH} \delta_{AH} e^{-\alpha_{AH} R_{AH}}\right), \tag{3.80}$$

where A ≠ H, C, N, O, and F.

The set of atoms for which AM1* parameters currently exist include P, S, Cl, Al, Si, Ti, Zr, Cu, Zn, Br, I, V, Cr, Co, Ni, Mn, Fe, Pd, and Ag,[31-39] along with their corresponding bond specific parameters.

3.2.7 AM1/d-PhoT

Similar to AM1*, Nam et al.[40] developed AM1/d-PhoT around the original AM1 Hamiltonian. For phosphorus the method includes *d*-orbitals to accurately model the hypervalent nature of the atom. In addition the method employs a modification to the AM1 core-core repulsion term, which is given as

$$E_{AB}^{AM1/d-PhoT} = E_{AB}^{MNDO} + \frac{Z_A Z_B}{R_{AB}} G_{scale}^A G_{scale}^B \left[\sum_i a_{iA} e^{-b_{iA}(R_{AB} - c_{iA})^2} + \sum_i a_{iB} e^{-b_{iB}(R_{AB} - c_{iB})^2} \right], \tag{3.81}$$

where all terms are defined exactly as those of the original AM1 core-core repulsion (eq. 3.70) and the two additional terms G_{scale}^A and G_{scale}^B are scaling parameters for atom A and B.

For AM1/d-PhoT the scaling parameters vary from 0 to 1, with values of 0 recovering the original MNDO core-core repulsion, while values of 1 recover the AM1 core-core repulsion.

Alternatively, the product $G_{scale}^{A} G_{scale}^{B}$ can be made into pairwise terms for specific atom pairs. The scaling parameters also provide the flexibility to attenuate (or even switch off) Gaussian core-core interactions between certain atoms and offers a simple mechanism for interconverting between AM1-like models and MNDO-like models.[40] Unlike AM1* (which keeps the original AM1 parameters for hydrogen, carbon, nitrogen, oxygen, and fluorine), the AM1/d-PhoT Hamiltonian only maintains the original carbon parameters of AM1, while the parameters of hydrogen, oxygen, and phosphorus were re-optimized.

3.2.8 PM6

Due to the increased accuracy achieved by Voityuk and Rösch[30] with the inclusion of diatomic parameters into the core-core repulsion (eq. 3.78), Stewart decided to include these parameters into the core-core interaction of PM6.[41] However, instead of the core-repulsion function converging to the exact point-charge interaction at increased interatomic distances (as is the case for AM1/d), PM6 makes use of a small perturbation to the core-core repulsion term resulting in an increased accuracy, especially for rare gas interactions.[41] The resulting *general* form of the core-core repulsion for PM6 is given as,

$$E_{AB}^{PM6} = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle \left(1 + \delta_{AB} e^{-\alpha_{AB}\left(R_{AB} + 0.0003\, R_{AB}^{6}\right)}\right). \tag{3.82}$$

For small distances the PM6 core-core repulsion is very similar to that of AM1/d. However, for distances longer than approximately 3 Å the PM6 core-repulsion function becomes significantly smaller than that of AM1/d.

During parameter optimization it was discovered that the calculated hydrogen bond interactions were too small and to correct for this the core-repulsion function was modified only for C-H and O-H interactions resulting in

$$E_{AH}^{PM6} = Z_A Z_H \langle s_A s_A \mid s_H s_H \rangle \left(1 + \delta_{AH} e^{-\alpha_{AH} R_{AH}^{2}}\right), \tag{3.83}$$

where *A* represents carbon or oxygen.

The equation only becomes important at hydrogen bond distances around 2 Å. By a decrease in the value of the exponential term the hydrogen bond energy increases.

Another problem encountered during parameter optimization was that all compounds containing $-C\equiv C-$ groups were found to be about 10 kcal/mol too stable by using the general form of the PM6 core-core repulsion term (eq. 3.82). To overcome this problem the core-core repulsion was once again modified for C-C interactions producing

$$E_{AB}^{PM6} = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle \left(1 + \delta_{AB} e^{-\alpha_{AB}\left(R_{AB}+0.0003\,R_{AB}^6\right)} + 9.28 e^{-5.98\,R_{AB}}\right), \quad (3.84)$$

where A and B represent the different interacting carbon atoms.

A final correction was added during the testing phase of the optimized PM6 parameters in which Stewart discovered that Si-O interactions were slightly repulsive instead of being slightly bound. To rectify this error a final correction to the core-core repulsion term was added for Si-O interactions giving rise to

$$E_{AB}^{PM6} = Z_A Z_B \langle s_A s_A \mid s_B s_B \rangle \left(1 + \delta_{AB} e^{-\alpha_{AB}\left(R_{AB}+0.0003\,R_{AB}^6\right)} - 0.0007 e^{-(R_{AB}-2.9)^2}\right), \quad (3.85)$$

where A and B represent silicon and oxygen atoms.

In addition to the core-core repulsion terms, a set of *d*-orbitals were also added to PM6 for many of the main-group elements and transition metals. This resulted in a method that is generally better than or comparable to previously available methods for the main-group elements. However, a statistical analysis showed that a recent re-parameterization of AM1, namely RM1,[42] performed more accurately than PM6 and any of the other NDDO methods for organic compounds.[1] The performance of PM6 for heats of formation of common organic compounds is better than B3LYP and HF methods at the 6-31G(d) level.[1,41] However, geometries predicted at the PM6 level are somewhat worse, and for electronic properties such as, ionization potential and dipole moments, it performs significantly worse than B3LYP and HF.[41]

3.2.9 PM7

As mentioned in Chapter 2 (Section 2.2.19) PM7[43] was developed to tackle the known faults experienced by its predecessor, PM6. With this method Stewart replaced the two-center two-electron integral $\langle ss \mid ss \rangle$ with the following equation,

$$\langle ss \mid ss \rangle = \frac{1}{R} e^{-0.22(R-7)^2} + \left(1 - e^{-0.22(R-7)^2}\right) \left(R^2 + \frac{1}{4}\left(\frac{1}{G_A} + \frac{1}{G_B} \right)^2 \right)^{-1/2}, \qquad (3.86)$$

where G_A and G_B are one-center two-electron integrals for atoms A and B, respectively.

The above mentioned equation is only applicable to all interatomic separations, R, of less than 7.0 Å. A consequence of this modification was that the nuclear-nuclear and electron-nuclear terms also had to be modified in a similar manner. Changes of this type are necessary in order to satisfy the requirement that there must be no net attraction or repulsion between any two well-separated neutral atoms.[43] An additional modification was made to correct a spurious contribution to the energy of solids arising from hybrid orbitals or lone pairs of which two types exist: the *s-p* type, best exemplified by the lone pair in ammonia, and the *s-d* type, found in some transition metal complexes. The correction was made to decrease the value of the hybrid NDDO integrals in a manner similar to eq. 3.86, resulting in integrals that converge to zero with increasing distance faster than the original NDDO integrals.

With the advent of linear scaling techniques, SE methods have become useful for modeling large biochemical systems such as DNA and proteins, particularly enzymes. In all such systems intermolecular interactions, especially hydrogen bonding, play an essential role, so the failure of these methods to accurately reproduce intermolecular interactions seriously limits their applicability and casts doubt on the validity of any results obtained.[43] In order to rectify the matter researchers have proposed various post-SCF dispersion and hydrogen bonding corrections which shall be discussed in more detail in the section that follows.

Most SE methods have low accuracy in reproducing barrier heights for reactions. There are various possible causes for this, including:

i) The restricted basis set used in SE methods which might preclude the development of a method that could simultaneously model both ground and transition states.

ii) Subtle electronic phenomena might occur in the region of the transition state because of a lowered HOMO−LUMO gap.

iii) The almost complete absence of transition state systems in the parameterization training set might result in a lack of definition in that region of parameter space.

In an attempt to improve the accuracy of predicting barrier heights, a specific parameterization has been for adopted for PM7 (entitled PM7-TS). This approach can be justified on pragmatic grounds: a method for predicting barrier heights with increased accuracy is likely to be useful when modeling chemical reactions. The only methodological change required by this parameterization was to freeze all geometries (reactants, transition states, and products) at their optimized PM7 structures.

3.2.10 Dispersion and Hydrogen bonding

Noncovalent interactions are of fundamental importance for chemistry and molecular biology, but a theoretical description of these interactions is difficult, mainly because they are much weaker than covalent interactions and also because of the key role played by the London dispersion energy.[44] In order to address this issue Řezáč et al.[44] made use of an empirical based dispersion correction for the PM6 Hamiltonian (entitled PM6-DH), in which the total dispersion energy is calculated as,

$$E_{disp} = -\sum_{ij} f_{damp}\left(r_{ij}, R_{ij}^{0}\right) C_{6ij} r_{ij}^{-6} , \tag{3.87}$$

where r_{ij} is the interatomic distance and R_{ij}^{0} is the equilibrium van der Waals (vdW) separation derived from the atomic vdW radii.[45] C_6 is a set of atomic dispersion coefficients that were acquired from work by Grimme,[46] in which each element is assigned only one dispersion coefficient. f_{damp} is a damping function that is present due to the fact that the r^{-6} is only an asymptotic expansion, i.e. it is not valid at short distances.

In addition to the dispersion based interactions another facet that is of utmost importance in various chemical systems is that of hydrogen bonding, which, like dispersion, is difficult to describe theoretically. Together with the development of the dispersion based correction Řezáč et al.[44] also introduced an H-bonding correction to the PM6 Hamiltonian that affected only hydrogen bonds. This correction had the following form,

$$E_{HB} = a\left[\frac{q_A q_H}{r^2} \times \cos(\theta) + bc^r\right], \tag{3.88}$$

where r is the A···H distance. θ denotes the angle between the donor (D) and acceptor (A) atoms (i.e. angle A···H−D). q_A and q_H are the charges on atoms A and H, respectively. a, b, and c are parameters fitted to obtain the best results over the training set (S22 dataset).[47]

PM6-DH did overestimate dispersion effects in saturated systems, as a result Korth et al.[48] established a new set of dispersion based parameters together with a more comprehensive H-bond correction (PM6-DH2)[48] given as,

$$E_{HB} = \left[a \times \frac{q_A q_H}{r^b} + c \times d^r\right] \times \cos(\theta) \times \cos(\phi) \times \cos(\psi), \tag{3.89}$$

with ϕ as the deviation of the R_2−A···H angle (R_2 is the donor "base atom") from the idealized optimal H-bond angle (taken as 109.48° for sp^3 and 120° for sp^2 structures) and ψ as the deviation of the R_1R_2A···H torsion angle from the idealized optimal H-bond torsion angle (taken as 109.48° for sp^3 hybridized nitrogen, 109.48° or 109.2° for other sp^3 hybridized structures, and 0° for sp^2 structures).

DH2 had the problem of being directly dependent on the distance between the hydrogen and the acceptor atom, resulting in the development of DH_+.[49] Here the H-bond correction is taken as a charge-independent atom-atom term between two atoms capable of serving as acceptor or donor (e.g. O, N). This term is then weighted by a function that accounts for the sterical arrangement of the two fragments relative to each other and the positioning of an H atom somewhere between them. This results in a correction with the following form,

$$E_{HB} = \frac{C_{AB}}{r_{AB}^2} \bullet f_{geom} \times f_{damp}, \tag{3.90}$$

where A and B are the two possible acceptor/donor atoms, C_A and C_B are hydrogen bond correction parameters, ϕ and ψ are symmetrically used for both the donor and acceptor atoms, f_{damp} is chosen so that no fitting is necessary (albeit the long range cutoff could in principle be taken as a fit parameter, e.g. if it turns out that the structures of very large molecules are found to be too dense) and is switched on between a donor−acceptor distance of 2.3 and 2.5 Å and slowly

switched off between 3.5 and 10.5 Å. The f_{bond} function brings the correction to zero if the hydrogen wanders away too far from both electronegative atoms (with r_{XH} being the smaller one of the two distances r_{AH} and r_{BH}), and this is switched off between 1.15 and 1.25 Å.

DH_+ also had some disadvantages (provided in Chapter 2, section 2.2.21), which Řezáč and Hobza[50] wished to address and improve upon with the D3H4 correction. With this correction the dispersion is modified from that given in eq. 3.87,

$$E_{disp} = s_6 \sum_{ij} f_{damp}\left(r_{ij}, R_{ij}^0\right) C_{6ij} r_{ij}^{-6}, \tag{3.91}$$

where s_6 is a correction term, and f_{damp} is a function damping the dispersion at short distances.

However, the dispersion correction given in eq. 3.91 does not yield satisfactory results for SE methods.[50] A specific error was encountered in the description of hydrocarbons where the intermolecular distance is strongly underestimated owing to weak Pauli repulsion between hydrogens. This cannot be corrected by the dispersion, which is only attractive. As a result D3H4 possess an additional repulsive term on all pairs of hydrogen atoms,

$$E_{rep} = s_{HH} \times \left(1 - \frac{1}{1 + \exp\left(-e_{HH} \left(\frac{r_{ij}}{R_{HH}^0} - 1 \right) \right)} \right), \tag{3.92}$$

where s_{HH} sets the strength of the correction, R_{HH}^0 determines the distance where the function acts, and the exponent e_{HH} determines how steep it is.

While the above mentioned repulsive correction is independent of the dispersion, in practical implementation it is calculated along with the dispersion correction. In addition to utilizing a new dispersion correction D3H4 also possess a modified hydrogen bond correction, which is given as,

$$E_{HB} = c \times f_{rad}\left(r_{DA}\right) \times f_{ang}\left(\alpha_{DHA}\right) \times f_{PT}\left(r_{DH}, r_{AH}\right) \times f_{charge} \times f_{wat}, \tag{3.93}$$

where *c* is the parameter determining the strength of the correction, α_{DHA} is the donor-hydrogen-acceptor angle (defined as zero in the linear arrangement), r_{DH} and r_{AH} are the distances between the hydrogen and the donor and acceptor, respectively. f_{rad} is the radial part that determines the strength of the correction from the donor−acceptor distance (r_{DA}) scaled by the angular term (f_{ang}), and the proton transfer term (f_{PT}), which depends on the position of the hydrogen between the donor and acceptor. f_{charge} is an additional scaling term that is applied for charged groups in order to make the correction stronger. f_{wat} is a scaling term that is applied in the case of water acting as the hydrogen donor.

The expansion of each of the terms provided in eq. 3.93 can be obtained from the original D3H4 paper[50] and shall *not* be discussed here as this correction was *not* utilized in this thesis. Among the tested methods, PM6-D3H4, DFTB-D3H4, and RM1-D3H4 yield errors lower than 1 kcal/mol in multiple benchmark datasets. In addition the methods reproduce geometries of non-covalent complexes with good accuracy, which makes them useful for many applications.[50]

Other variants of dispersion and hydrogen bonding corrections do exist and some of these have been mentioned briefly in Chapter 2 (Section 2.2.21), but for purposes of this thesis the theory surrounding these methods shall *not* be discussed.

3.3 Density Functional Theory

The central idea underpinning density functional theory (DFT) is that there is a relationship between the total electronic energy and the overall electronic density.[51] This idea was established by Hohenberg and Kohn[52] who showed that the ground-state energy and other properties of a system were uniquely defined by the electron density. The authors established this relationship by means of a functional given as,

$$E[\rho(r)] = \int V_{ext}(r)\rho(r)dr + F[\rho(r)], \tag{3.94}$$

where the energy *E* depends on a function of the electron density $F[\rho(r)]$, which is a function of the nuclear and electronic coordinates *r*. The external potential $V_{ext}(r)$ is the Coulomb interaction with the nuclei. $F[\rho(r)]$ is further expanded into the sum of kinetic energy of the electrons and inter-electronic contributions,

$$F[\rho(r)] = E_{KE}[\rho(r)] + E_H[\rho(r)] + E_{XC}[\rho(r)], \tag{3.95}$$

where $E_{KE}[\rho(r)]$ is the kinetic energy, $E_H[\rho(r)]$ is the electron-electron repulsion energy, and $E_{XC}[\rho(r)]$ is the exchange and correlation contribution to the energy.[51]

The matrix solution to these equations takes the form of the Kohn-Sham equations, which are identical in form to the Roothaan-Hall equations (eq. 3.39),[51]

$$H^{KS}C = SC\varepsilon, \tag{3.96}$$

where the H^{KS} is the Kohn-Sham Hamiltonian. These equations are variational and self-consistent where an approximate density functional ρ_o is chosen and iteratively improved upon.[12]

The $E_{KE}[\rho(r)]$ and $E_H[\rho(r)]$ functionals given in eq 3.95 are chosen as follows,

$$E_{KE}[\rho(r)] = \sum_{i=1}^{N} \int \psi_i(r) \left(-\frac{\nabla^2}{2} \right) \psi_i(r) dr, \tag{3.97}$$

$$E_H[\rho(r)] = \frac{1}{2} \iint \frac{\rho(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2, \tag{3.98}$$

where $E_{KE}[\rho(r)]$ describes a system of non-interacting electrons and $E_H[\rho(r)]$ is the Hartree electrostatic energy that is the sum of all pairwise electrostatic interactions.

$E_{XC}[\rho(r)]$ is usually divided into exchange and correlation parts,

$$E_{XC}[\rho(r)] = E_X[\rho(r)] + E_C[\rho(r)], \tag{3.99}$$

where $E_X[\rho(r)]$ describes the exchange contribution which are due to same-spin interactions, and $E_C[\rho(r)]$ describes the correlation contribution due to mixed-spin interactions.

Functionals differ in the way they treat exchange and correlation. Local functionals are based on the electron spin densities (ρ), while gradient-corrected functionals depend on the electron spin densities as well as their gradient ($\nabla\rho$). Local density approximations (LDA) treat the density as a uniform electron gas, while the more advanced gradient methods treat the density as a non-uniform electron gas. Examples of LDAs are Perdew-Wang (PW), and Vosko, Wilk, Nusair

(VWN) functionals and gradient-corrected functionals are the Becke exchange; Lee, Yang, Parr (LYP) correlation.[12,53]

Becke's hybrid three parameter functional including LYP correlation (B3LYP)[54,55] linearly combines the Hartree-Fock exchange with linear and gradient-corrected exchange terms and correlation terms,

$$E_{XC}^{B3LYP} = (1-a_0)E_X^{LSDA} + a_o E_X^{HF} + a_X \Delta E_X^{B88} + a_C E_C^{LYP} + (1-a_C)E_C^{VWN}, \tag{3.100}$$

where the exchange part (all terms with E_x) is composed of local spin density approximated (LSDA) exchange, the Hartree-Fock exchange, and Becke's original exchange function (Becke-88 or B88). The correlational part (all terms with E_c) comprises the LYP and the VWN correlational functional. The empirically derived coefficients a_o, a_x, and a_c are 0.20, 0.72, and 0.81, respectively.[12,56]

3.4 SCC-DFTB

The self-consistent charge density functional tight-binding (SCC–DFTB) method involves a second-order expansion of the DFT total energy functional with respect to the charge density fluctuations around a given reference density, ρ_0 . The energy is expressed as,

$$E^{SCC} = \sum_{i\mu\nu} c_\mu^i c_\nu^i H_{\mu\nu}^0 + \frac{1}{2}\sum_{\alpha\beta}\gamma_{\alpha\beta}\Delta q_\alpha \Delta q_\beta + \frac{1}{2}\sum_{\alpha\beta} V[\rho_0^\alpha; p_0^\beta; R_{\alpha\beta}], \tag{3.101}$$

where $c_{\mu/\nu}^i$ are the linear combination of atomic orbital (LCAO) coefficients, $H_{\mu\nu}^0$ the Hamiltonian matrix with reference density, $\Delta q_{\alpha/\beta}$ the Mulliken charges on atom α/β and $V[\rho_0^\alpha; p_0^\beta; R_{\alpha\beta}]$ the repulsive potential.

In the original version of SCC–DFTB the $\gamma_{\alpha\beta}$ function was approximated as,

$$\gamma_{\alpha\beta} = \frac{1}{R_{\alpha\beta}} - S_{\alpha\beta}, \tag{3.102}$$

where $S_{\alpha\beta}$ is an exponentially decaying short-range function that describes the deviation of the $\gamma_{\alpha\beta}$ function from the $\frac{1}{R_{\alpha\beta}}$ behavior with increasing overlap of α and β. It is also responsible for the convergence of $\gamma_{\alpha\beta}$ to a finite value at zero distance.

In 2001, Elstner et al.[57] included an empirical dispersion,

$$E = E^{SCC} - \sum_{\alpha\beta} f(R_{\alpha\beta}) \frac{C_6^{\alpha\beta}}{R_{\alpha\beta}^6}, \tag{3.103}$$

where $C_6^{\alpha\beta}$ is a van der Waals coefficient obtained from atomic polarizablities and for short distances the $\frac{1}{R^6}$ term should be damped, where the electronic densities start to overlap. The damping function is given by,

$$f(R_{\alpha\beta}) = \left[1 - \exp\left(-d * \left(\frac{R_{\alpha\beta}}{R_0^{\alpha\beta}} \right)^N \right) \right]^M , \tag{3.104}$$

where $d = 3.0$, $N = 7$, $M = 4$ and $R_0 = 3.8$Å for first row elements.

R_0 is defined by making use of the cubic mean rule,[58]

$$R_0^{\alpha\beta} = \frac{\left(R_0^\alpha\right)^3 + \left(R_0^\beta\right)^3}{\left(R_0^\alpha\right)^2 + \left(R_0^\beta\right)^2} . \tag{3.105}$$

In addition to the empirical dispersion mentioned above, Elstner[59] added a damping function to the $S_{\alpha\beta}$ term in the $\gamma_{\alpha\beta}$ function (eq. 3.102) for atomic pairs involving hydrogen atoms to improve hydrogen bonding interactions,[60]

$$\gamma_{\alpha H} = \frac{1}{R_{\alpha H}} - S_{\alpha H} \exp\left[-\left(\frac{U_\alpha + U_H}{2} \right)^\zeta R_{\alpha H}^2 \right], \tag{3.106}$$

where U_α is the atomic Hubbard parameter that is related to the chemical hardness of atom α.[61] ζ is a parameter that was adjusted by fitting G3B3[62] energies for small hydrogen bond clusters and MP2/G3large[63] energies for large complexes.

The theory provided here for SCC–DFTB is a brief overview of the method and for further details the reader is referred to the original papers[64,65] and reviews.[66-68] For the current work we make use of the methodology given above, as a result the more recent modifications made to the method (improvements for Coulomb interactions between atomic partial charges and complete

third-order expansion of the DFT total energy)[69] are outside the scope of this thesis and shall *not* be discussed further.

3.5 M06 and M06-2X

M06 and M06-2X are hybrid meta-generalized gradient approximation (GGA) density functionals.[70] The M06 functional form is a linear combination of the functional forms of M05[71,72] and VSXC[73] exchange functionals, given as,

$$E_X^{M06} = \sum_\sigma \int dr \left[F_{X\sigma}^{PBE}(\rho_\sigma, \nabla\rho_\sigma) f(w_\sigma) + \varepsilon_{X\sigma}^{LSDA} h_X(x_\sigma, z_\sigma) \right], \tag{3.107}$$

where,

$$h_X(x_\sigma, z_\sigma) = \left(\frac{d_0}{\gamma(x_\sigma, z_\sigma)} + \frac{d_1 x_\sigma^2 + d_2 z_\sigma}{\gamma_\sigma^2(x_\sigma, z_\sigma)} + \frac{d_3 x_\sigma^4 + d_4 x_\sigma^2 z_\sigma + d_5 z_\sigma^2}{\gamma_\sigma^3(x_\sigma, z_\sigma)} \right), \tag{3.108}$$

$F_{X\sigma}^{PBE}(\rho_\sigma, \nabla\rho_\sigma)$ is the exchange energy density of the PBE[74] exchange model, $\varepsilon_{X\sigma}^{LSDA}$ is the local spin density approximation for exchange.[75] $f(w_\sigma)$ is the spin kinetic-energy-density enhanced factor,

$$f(w_\sigma) = \sum_{i=0}^{m} a_i w_\sigma^i , \tag{3.109}$$

where the variable w_σ is a function of t_σ, which in turn is a function of the spin kinetic energy density τ_σ and spin density ρ_σ,

$$w_\sigma = (t_\sigma - 1) \Big/ (t_\sigma + 1), \tag{3.110}$$

$$t_\sigma = \tau_\sigma^{LSDA} \Big/ \tau_\sigma . \tag{3.111}$$

The exchange functional in M06-2X is a special case in which $h_X(x_\sigma, z_\sigma) = 0$. For this the M06 functional form for exchange merely reduces to the M05 functional form.

The functional form of the M06 and M06-2X correlation functionals is the same as the functional form of the M06-L[76] or M06-HF,[77] in which opposite-spin and parallel-spin correlation are treated differently.[70] Opposite-spin correlation energy is expressed as,

$$E_C^{\alpha\beta} = \int e_{\alpha\beta}^{UEG}\left[g_{\alpha\beta}\left(x_\alpha, x_\beta\right) + h_{\alpha\beta}\left(x_{\alpha\beta}, z_{\alpha\beta}\right)\right] dr, \tag{3.112}$$

where $g_{\alpha\beta}\left(x_\alpha, x_\beta\right)$ is defined as,

$$g_{\alpha\beta}\left(x_\alpha, x_\beta\right) = \sum_{i=0}^{n} cC_{\alpha\beta,i}\left(\frac{\gamma C_{\alpha\beta}\left(x_\alpha^2 + x_\beta^2\right)}{1 + \gamma C_{\alpha\beta}\left(x_\alpha^2 + x_\beta^2\right)}\right)^i, \tag{3.113}$$

and $h_{\alpha\beta}\left(x_{\alpha\beta}, z_{\alpha\beta}\right)$ is defined in eq. 3.108, with $x_{\alpha\beta}^2 = x_\alpha^2 + x_\beta^2$ and $z_{\alpha\beta}^2 = z_\alpha^2 + z_\beta^2$.

Parallel spin correlation energy is expressed as,

$$E_C^{\sigma\sigma} = \int e_{\sigma\sigma}^{UEG}\left[g_{\sigma\sigma}\left(x_\sigma\right) + h_{\sigma\sigma}\left(x_\sigma, z_\sigma\right)\right] D_\sigma dr, \tag{3.114}$$

with $g_{\sigma\sigma}\left(x_\sigma\right)$ given as,

$$g_{\sigma\sigma}\left(x_\sigma\right) = \sum_{i=0}^{n} cC_{\sigma\sigma,i}\left(\frac{\gamma C_{\sigma\sigma} x_\sigma^2}{1 + \gamma C_{\sigma\sigma} x_\sigma^2}\right)^i, \tag{3.115}$$

and $h_{\sigma\sigma}\left(x_\sigma, z_\sigma\right)$ is defined in eq. 3.108. D_σ is the self-interaction correction factor,

$$D_\sigma = 1 - \frac{x_\sigma^2}{4\left(z_\sigma + C_F\right)}. \tag{3.116}$$

From the above mentioned equations the total M06 correlation energy is given as,

$$E_C = E_C^{\alpha\beta} + E_C^{\alpha\alpha} + E_C^{\beta\beta}. \tag{3.117}$$

The hybrid exchange-correlation energy for M06 can be written as,

$$E_{XC}^{hyb} = \frac{X}{100} E_X^{HF} + \left(1 - \frac{X}{100}\right) E_X^{DFT} + E_C^{DFT}, \tag{3.118}$$

where E_X^{HF} is the nonlocal Hartree-Fock (HF) exchange energy, X is the percentage of HF exchange in the hybrid functional, E_X^{DFT} is the local DFT exchange energy and E_C^{DFT} is the local DFT correlation energy.

The parameter X along with the parameters in the new exchange (eqs. 3.107−3.109) and correlation functionals (eqs. 3.111−3.114) were optimized during the development of M06 and M06-2X.[70] This produced a set of functionals that are well suited for the study of organometallic, inorganometallic and main-group thermochemistry. In addition the functionals are recommended for main-group kinetics and non-covalent interactions.[70] Due to this and the inherent success of the functional,[78-80] M06-2X was used for all DFT based simulations conducted in this work.

3.6 References

(1) Kayi, H. PhD Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.
(2) Born, M.; Oppenheimer, J. R. *Ann. Phys. (Leipzig)* **1927**, *84*, 457.
(3) Sherrill, C. D. *An Introduction to Hartree-Fock Molecular Orbital Theory*, School of Chemistry and Biochemistry, Georgia Institute of Technology, 2000.
(4) Pauli, W. *Z. Physik* **1925**, *31*, 765.
(5) Fock, V. *Z. Physik* **1930**, *61*, 126.
(6) Slater, J. C. *Phys. Rev.* **1929**, *34*, 1293.
(7) Slater, J. C. *Phys. Rev.* **1930**, *35*, 509.
(8) Hückel, E. *Z. Physik* **1931**, *70*, 204.
(9) Hückel, E. *Z. Physik* **1931**, *72*, 310.
(10) Hückel, E. *Z. Physik* **1932**, *76*, 628.
(11) Cramer, C. J. *Essentials of Computational Chemistry*; 2nd ed.; John Wiley and Sons, Ltd., 2004.
(12) Jensen, F. *Introduction to Computational Chemistry*; 2nd ed.; John Wiley and Sons, Ltd., 2007.
(13) Levine, I. A. *Quantum Chemistry*; 5th ed.; Prentice Hall, Upper Saddle River, New Jersey 07458, 2000.
(14) Lewars, E. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*; 2nd ed.; Kluwer Academic Publishers, 2003.
(15) Roothan, C. C. J. *Rev. Mod. Phys.* **1951**, *23*, 69.
(16) Hall, G. G. *Proc. Royal Soc. London A* **1951**, *205*, 541.
(17) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(18) Stewart, J. J. P. *J. Comp. Aid. Mol. Des.* **1990**, *4*, 1.
(19) Oleari, L.; DiSipio, L.; DeMichelis, G. *Mol. Phys.* **1966**, *10*, 97.
(20) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(21) Burstein, K. Y.; Isaev, A. N. *Theor. Chim. Acta* **1984**, *64*, 397.

(22) Winget, P.; Selçuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. *Theor. Chem. Acc.* **2003**, *110*, 254.
(23) Fiedler, L.; Gao, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2011**, *7*, 852.
(24) Zhang, P.; Fiedler, L.; Leverentz, H. R.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**, *7*, 857.
(25) Stewart, J. J. P. *J Comput Chem* **1989**, *10*, 221.
(26) Stewart, J. J. P. *J Comput Chem* **1989**, *10*, 209.
(27) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391.
(28) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, *93*, 315.
(29) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616.
(30) Voityuk, A. A.; Rösch, N. *J. Phys. Chem. A* **2000**, *104*, 4089.
(31) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, *9*, 408.
(32) Winget, P.; Clark, T. *J. Mol. Model.* **2005**, *11*, 439.
(33) Kayi, H.; Clark, T. *J. Mol. Model.* **2007**, *13*, 965.
(34) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 295.
(35) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 1253.
(36) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 29.
(37) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 1109.
(38) Kayi, H.; Clark, T. *J. Mol. Model.* **2011**, *17*, 2585.
(39) Kayi, H. *J. Mol. Model.* **2010**, *16*, 1029.
(40) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(41) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(42) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
(43) Stewart, J. J. P. *J Mol Model* **2013**, *19*, 1.
(44) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749.
(45) Bondi, A. *J. Chem. Phys.* **1964**, *68*, 441.
(46) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
(47) Jurečka, P.; Sponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
(48) Korth, M.; Pitonak, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344.
(49) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808.
(50) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141.
(51) Leach, A. R. *Molecular Modelling: Principles and applications*; 2nd ed.; Prentice Hall, 2001.
(52) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *B136*, 864.
(53) Jensen, J. H. *Molecular Modeling Basics*; Taylor & Francis Group, 2010.
(54) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(55) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785.
(56) Barnett, C. B. PhD Thesis, University of Cape Town, 2010.
(57) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
(58) Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827.
(59) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.
(60) Choi, T. H.; Jordan, K. D. *J. Phys. Chem. B* **2010**, *114*, 6932.
(61) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512.

(62) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.
(63) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
(64) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
(65) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.
(66) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.
(67) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.
(68) Elstner, M.; Frauenheim, T.; Suhai, S. *J. Mol. Struct. (Theochem)* **2003**, *632*, 29.
(69) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931.
(70) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
(71) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2006**, *2*, 364.
(72) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2005**, *123*, 161103.
(73) Van Voorhis, T.; Scuseria, G. E. *J. Chem. Phys.* **1998**, *109*, 400.
(74) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(75) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.
(76) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
(77) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
(78) Walker, M.; Harvey, A. J. A.; Sen, A.; Dessent, C. E. H. *J. Phys. Chem. A* **2013**, *117*, 12590.
(79) Peverati, R.; Truhlar, D. G. *Phil. Trans. R. Soc. A* **2014**, *372*.
(80) Brás, N. F.; Perez, M. A. S.; Fernandes, P. A.; Silva, P. J.; Ramos, M. J. *J. Chem. Theory Comput.* **2011**, *7*, 3898.

4. Semi-empirical parameterization

Both the procedure and algorithm (genetic algorithm) used for the parameterization of the two newly developed semi-empirical methods (AM1/d-CB1 and AM1-CB1) are discussed.*

4.1 Introduction

Parameterization is a very important part of the development of a new method and the accuracy of the method is heavily dependent on the quality of the parameters generated.[1] When conducting a semi-empirical (SE) parameterization a number of steps need to be taken into account, these include:

i) Deciding on an appropriate SE methodology to make use of as a starting point.
ii) Defining a training dataset which shall be used during the parameterization.
iii) Defining the properties which should be utilized during the parameterization process.
iv) Defining an appropriate error function (fitness) which upon optimization will produce the best set of parameters.
v) Deciding upon a mathematical algorithm which shall iteratively generate parameters and stop the process once an optimum set of parameters has been obtained (usually once a particular fitness has been achieved).

In the sections that follow we shall address each of the points mentioned above in a bit more detail.

4.2 Methodology

As mentioned in Chapter 1 the goal of this work is to model chemical glycobiological problems of biochemical interest. In order to do so a method which can model a combination of both organic and hypervalent systems needs to be chosen. In this work two such methods were selected, namely AM1/d-PhoT[2] and AM1*.[3] Despite both methods vast applicability in various chemical regimes,[2,4-12] the chapters that follow shall show that these methods perform very poorly when trying to reproduce key metrics in chemical glycobiology (i.e., sugar ring pucker, phosphate participation in transferase reactions). As a result a re-parameterization of both

methods was sought after with the aim of producing holistic biochemical QM/MM toolsets able to simulate fundamental problems of binding and enzyme reactivity in chemical glycobiology.

4.3 Training dataset

The training dataset used in parameterization is extremely important as it governs the generation of an accurate set of parameters. This dataset should reflect the main characteristics of the chemical class of systems for which it is designed. Thus, the data must be as accurate as possible and it must represent a wide range of chemical systems and properties.[1] It must also be suitable to be manipulated by mathematical tools permitting optimization of parameters. Such data can either be experimentally determined or obtained from high-level (*ab initio* or DFT) calculations. It is also possible, as in the current work, that the data is a combination of both experimental values and high-level calculation results. Using the approach of Dewar,[13] in which only experimentally determined data was used for the training set, has the advantage of circumventing any theoretical inadequacy of high-level methods. Parameterizing with only high-level reference values makes the SE method vulnerable to the errors experienced by the high-level method. Unfortunately, experimental data is not always available and in some cases the data is unreliable, making the use of high-level simulations a necessity. In this work we are confident that the high-level data utilized is accurate and reliable enough to challenge the experimental results.

The training dataset used in this thesis comprised of various molecules and molecular fragments important in chemical glycobiology, these include:

i) Various α- and β- conformers of glucopyranose and ribofuranose, which are the basic building blocks of glycobiological systems.

ii) Various α- and β- conformers of glucose-6-phosphate and ribose-5-phosphate, which are important in biologically significant systems such as DNA and RNA.

iii) Some important amino acids, such as; aspartic acid, asparagine, glutamic acid, histidine, argenine, phenylalanine, tyrosine, and tryptophan. These play key roles in catalytic based glycobiological reactions.

iv) Some general organic molecules, such as; H_2O, CH_3OH, C_2H_5OH, C_6H_5OH, CH_3CO_2H, CH_3OCH_3, $P(CH_3)_3$, $(CH_3)_3PO$, and H_3PO_4.

v) Fragments of glucopyranose and ribofuranose (1, 2-ethanediol and methoxymethanol).

Experimental data for various molecules in the training set were obtained from various sources, including: the QCRNA database,[14] the NIST chemistry webbook,[15] and various published papers.[2,16,17] The starting structures for the molecular systems were obtained from work by Barnett and Naidoo (β-D-glucopyranose),[18,19] Jalbout et al. (β-D-ribofuranose),[20] the QCRNA database (organic and phosphoric systems),[14,21] and the training dataset used in the development of the SE PM6 Hamiltonian (amino acids).[22] The α- conformers of the carbohydrates were constructed from their DFT optimized β- counterparts prior to undergoing an optimization themselves. All DFT based simulations were conducted with the M06-2X functional[23] together with the 6-311++G(3df, 2p) basis set. The functional was chosen due to the fact that it was specifically parameterized for non-metals and was recommended for use on main-group element systems, producing good thermochemistry, kinetics, and non-covalent interactions.[23] No energy refinement was necessary as a sufficiently large basis set was utilized during the geometry optimization process. Frequency calculations were conducted on all optimized structures to verify the nature of all stationary points and obtain thermochemical data. The DFT simulations were all conducted with the software package Gaussian 09.[24]

4.4 Properties

The properties chosen for the current work are heats of formation (ΔH_f), dipole moments, ionization potentials, proton affinities, interaction energies, and ring flexibility. The properties and their definitions are described in the sections that follow.

4.4.1 Heat of formation

This is obtained when the energy required to ionize the valence electrons of the atoms involved, $E_{isol}(A)$ (calculated using SE parameters), and heat of atomization, $\Delta H_f(A)$, are added to the electronic and core-core (nuclear) energy terms described in Chapter 3, yielding,

$$\Delta H_f = E_{el} + E_{AB}^{CORE} + \sum_A E_{isol}(A) + \sum_A \Delta H_f(A), \quad (4.1)$$

4.4.2 Dipole moment

The dipole moment (μ) is a measure of net molecular polarity, which is the magnitude of the charge (Q) at either end of the molecular dipole times the distance (r) between the charges,

$$\mu = Qr\,, \tag{4.2}$$

4.4.3 Ionization potential

The ionization potential (IP), or ionization energy, of a molecule refers to the *minimum* energy needed to remove an electron from its ground state to infinity, i.e. to form a radical cation of a species. The IP of a "stable" species (molecule that possess a relative minimum on a potential energy surface, is always positive.[25] There are two types of IPs, namely *vertical* and *adiabatic*. The *vertical* IP is produced when the energy difference between the precursor molecule M_1 and the species M_2, formed by removing an electron, have the same molecular geometry. The *adiabatic* IP arises when M_2 has a geometry that differs from that of M_1, i.e. M_2 has its own equilibrium geometry.

4.4.4 Proton affinity

The protonation states of, pK_a values, of biological systems plays an important role in both structure and reactivity. Considerable effort has been devoted to the prediction of proton affinities (PAs) and pK_a values with quantum chemistry.[26-36] However, this is an area that remains challenging due to the small differences in free energy that give rise to pK_a shifts (1 pK_a unit = 1.364 kcal/mol at 298.15K). On the other hand reliable prediction of pK_a shifts using known experimental data may not always be possible since the determination of experimental pK_a values of an appropriate reference state may not be available.

4.4.5 Interaction energy

The interaction energy is the contribution to the total energy that is caused by an interaction between the molecules being considered. It is calculated by taking the difference between the energies of isolated molecules (monomers) and their interacting assembly (dimers),

$$\Delta E_{\text{int}} = E(A,B) - \big(E(A) + E(B)\big), \tag{4.3}$$

where *E(A)* and *E(B)* are energies of the isolated molecules and *E(A, B)* is the energy of the interacting assembly. With SE methods these energies correspond to the heat of formation of the monomers and dimers, respectively.

4.4.6 Ring flexibility

A property which is of considerable importance in glycobiology is that of carbohydrate ring puckering. The carbohydrate ring pucker can be described by making use of a method called triangular decomposition/tessellation[37] which has been discussed, in detail, in Chapter 1. A few gas phase quantum mechanical (QM) molecular dynamics (100 ps) simulations were conducted on tetrahydrofuran and β-D-glucopyranose in which the QM region was treated with SCC-DFTB.[38] In addition Langevin dynamics at 298.15 K and group based cutoffs of 16.0, 14.0, and 12.0 Å were used. It would have been more accurate to treat the QM region with an *ab inito* based method, but this would require an extensive amount of compute time. SCC-DFTB was chosen because it is considerably faster than standard DFT based methods and it has shown to provide an accurate description of carbohydrate ring pucker.[19] Although it would have been preferable to model the 5-membered ring flexibility with β-D-ribofuranose, we discovered that gas phase QM (SCC-DFTB) MD simulations of this species yield strong hydrogen bonds between the hydroxyl located at the anomeric carbon (C1) and the primary alcohol at C4 (Figure 4.1) resulting in two indistinguishable pucker angles. Due to the two different substituents on C1 (hydroxyl group) and C4 (primary alcohol) it is expected that the ring would pucker at different rates relative to the reference plane (Chapter 1, Figure 1.6). With that not being the case we decided that it would be inaccurate to represent the 5-membered ring pucker by gas phase SCC-DFTB simulations of β-D-ribofuranose. As a result the five membered ring pucker was represented by tetrahydrofuran.

Using the pucker definitions described in Chapter 1 (eq. 1.1), together with the data generated from the MD simulations, a number of time correlation functions where generated for the pucker angles of tetrahydrofuran (θ_0 and θ_1) and β-D-glucopyranose (θ_0, θ_1, and θ_2). From the resulting correlation functions 30 ps of data were fitted using an exponential function given as,

$$f(x) = a \times \exp^{\left(-x/\tau\right)} + b\,, \tag{4.4}$$

where τ is the relaxation time.

The above mentioned τ value is then used during the parameterization process. To the best of our knowledge this is the *first* time that such a property is utilized during a SE parameterization.

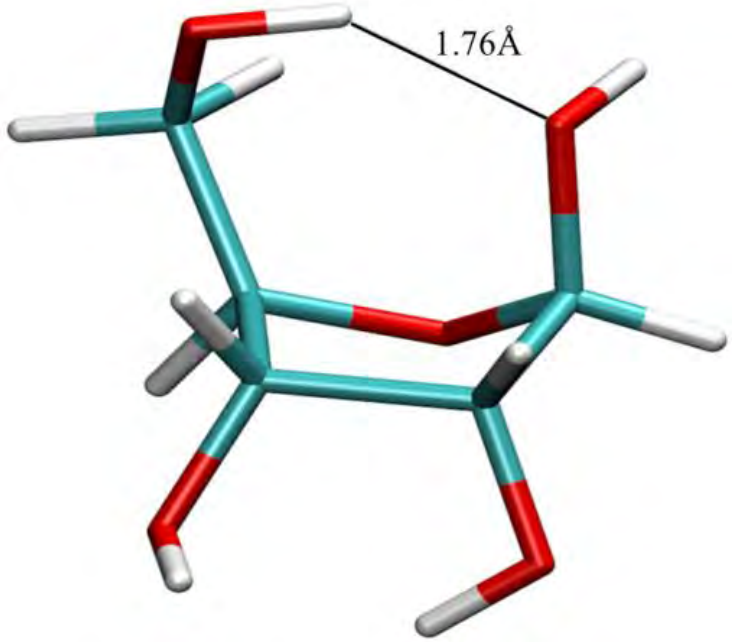


Figure 4.1: Structure of β-D-ribofuranose depicting the predominant hydrogen bond interaction present during a gas phase QM (SCC-DFTB) MD simulation.

4.5 Fitness

The fitness (or error function) is a function which upon minimization will yield the best set of SE parameters. This function plays crucial role during parameter optimization. The error function was defined as,

$$S = \sum_{i}^{mol} \sum_{j}^{prop(i)} w_{ij} \left(x_{ij}^{calc} - x_{ij}^{ref} \right)^2 , \tag{4.5}$$

where the first summation with the index i run over all molecules and the second summation with the index j runs over the properties associated with the ith molecule. w_{ij} is a weighting factor, x_{ij}^{calc} is the calculated molecular property j for molecule i using the generated set of parameters, and x_{ij}^{ref} is the corresponding target value (either M06-2X/6-311++G(3df, 2p), SCC-DFTB, or experimental data).

All calculated properties where computed with an in-house version of MOPAC, entitled MOPAC7.2,[39] as well as CHARMM35/MNDO97[40,41] interface. It is worth noting that a method such as AM1/d-PhoT only existed within the CHARMM/MNDO97 interface, but in this work we successfully transferred this Hamiltonian into MOPAC7.2. In addition, the AM1* Hamiltonian exists in the commercial software package VAMP,[42] as well as the software package EMPIRE,[43] which is freely available to academics. In the current work we have succeeded in coding the AM1* Hamiltonian into the CHARMM35/MNDO97 interface (permitting the use of hybrid QM/MM simulations with AM1*) as well as into the in-house MOPAC7.2. All SE simulations conducted with MOPAC7.2 were run as single point calculations on the DFT optimized geometries to ensure that structures such as half-chairs of β-D-glucopyranose, for example, remain half-chairs and do not produce boat conformers, as would be the case during geometry optimization. Due to the simulations only being single point *no* geometrical properties, such as bond lengths, bond angles, and dihedral angles, needed to be included in the parameterization. As shown in eq. 4.2 each property utilized in the fitness has a weight associated with it and the weights used in this work are provided in Table 4.1. These weights were chosen to render various properties unit less and to increase the significance of certain properties over others during function evaluation.

Table 4.1: Weighting factors used for reference properties

Reference data	Weight	Unit
ΔH_f	1.0	$mol.kcal^{-1}$
Dipole moments	25.0	$Debye^{-1}$
Ionization potential	15.0	eV^{-1}
Proton affinities	25.0	$mol.kcal^{-1}$
Interaction energies	20.0	$mol.kcal^{-1}$
Correlation time	15.0	s^{-1}

Various weighting factors were tested during parameter optimization and these were found to produce optimum results (lowest fitness).

4.6 Parameter optimization

The error function provided above was evaluated with the aid of an algorithm which is typically utilized in evolutionary computing. This particular algorithm is known as a genetic algorithm (GA) and is based on the original work of Goldberg.[44] It was inspired by Darwin's

theory of evolution, which states that the survival of an organism is affected by the rule "the strongest species that survives". GA's are stochastic global search algorithms with heuristic ideas (operators) borrowed from the mechanisms of natural selection and natural genetics. Although computational simple, these algorithms are particularly powerful in their search for a minimum. Furthermore, they are not fundamentally limited by restrictive assumptions about the search space. The search is not exact meaning that there is no guarantee that the global minimum will be found, but the result typically is a very low-valued local minimum.[45]

In a GA evolution starts from a population of randomly generated individuals, which is an iterative process. The population in each iteration is called a generation. Within each generation the fitness of every individual in the population is evaluated, were the fitness is provided in eq. 4.2. The more *fit* individuals are stochastically selected from the current population, and each individual is modified via a crossover and/or mutation to form a new generation. The new generation of candidate solutions is then used in the next iteration of the GA. This process continues until either an appropriate fitness has been achieved or the maximum number of generations has been reached. It is worth mentioning that as far as mutation is concerned one should never choose too small or too large a mutation rate. Too small a mutation may lead to genetic drift (non-ergodic in nature), and too large a mutation may lead to loss of good solutions, unless there is elitist selection in which better parameters from the current generation carry over to the next, unaltered. A flowchart illustrating the mechanics of a GA is provided in Figure 4.2. This algorithm was chosen in the current work since it has shown significant promise when establishing SE based parameter sets,[1,2,46,47] and it is a method which can sample a much wider range of parameter space than most algorithms.

The parameterization procedure followed in this work was three-fold:

i) Obtain a parameter set for H, C, and O which will accurately model the ring puckering of the carbohydrates.

ii) Optimize the parameters of N and P while fixing those of H, C, and O to the values obtained above, in order to provide a better energetic description for the amino acids and phosphates present in the training set.

iii) Do a final refinement of all H, C, N, O, and P parameters.

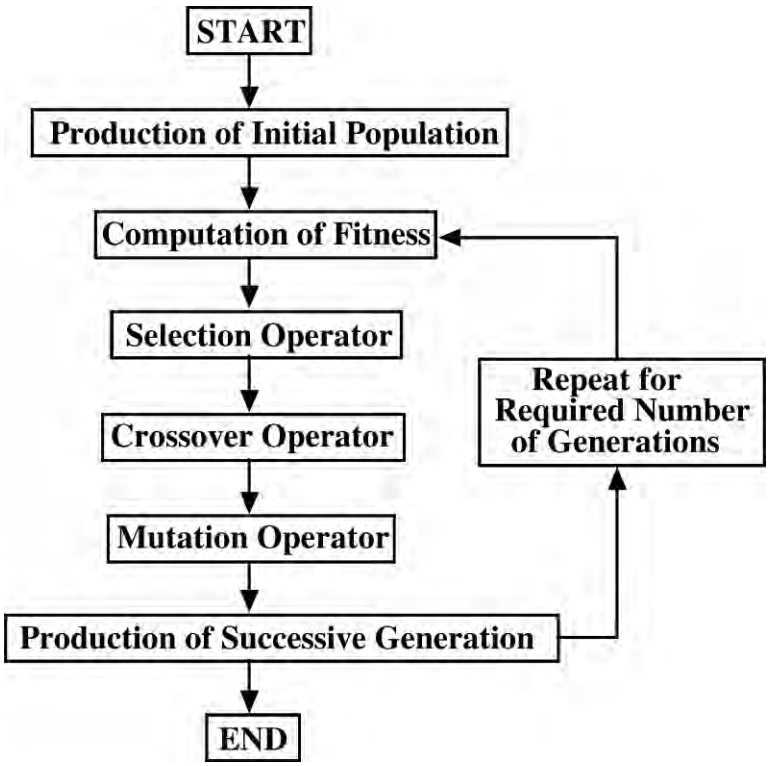


Figure 4.2: Flowchart of a genetic algorithm.

The success of the GA, or any algorithm for that matter, depends on the initial starting point used for the parameter optimization. In order to decide upon an appropriate set of starting parameters a number of simulations had to be run on the current training set with various SE methods (AM1, PM3, PMCARB-1, $PM3^{MS}$, RM1, AM1/d-PhoT, and AM1*), focusing specifically on the properties mentioned above. Results for these simulations, which shall be provided in the chapters that follow, show that in most cases it is the AM1 Hamiltonian which provides the best mean signed errors (MSE) and mean unsigned errors (MUE). However, AM1 has shown poor performance when applied to carbohydrate ring puckering.[19] As a result it was decided that use of standard AM1 parameters as a starting point is not a good idea. Having eliminated the AM1 parameter set, the other methods which produced small MSE and MUE were RM1, AM1/d-PhoT, and AM1*. RM1 generally produced good results for organic systems (ΔH_f MUE 4.8) and reasonable results for systems which possess phosphorus (ΔH_f MUE 15.1 kcal/mol). However, it is known that RM1 does not apply *d*-orbitals onto the hypervalent phosphorus. The only methods that do incorporate *d*-orbitals onto phosphorus are AM1/d-PhoT and AM1* and these methods would provide a better representation for the hypervalent species, accounting for the complex *d*-

orbital interactions that this atom is capable of having. Therefore the methods that were chosen as starting points for our re-parameterization were RM1 (H, C, N and O), AM1/d-PhoT (P) and AM1* (P).

The initial optimization of H, C, and O was conducted by making use of a GA. Due to gas phase MD simulations having to be run at every GA iteration the population size for the GA had to be small enough to be computationally efficient, while not too small so as to miss the minima along parameter space. After extensive testing a population size of 100 was used for our parameterization purposes. The mutation rate and elitism were set to their default values of 0.50 and 0.05, respectively. Initially only the parameters for H, C, and O were optimized while the parameters of N and P were fixed to their RM1 values and AM1/d-PhoT or AM1* values, respectively. The parameter sets with the lowest fitness (minimum value for *S* of eq. 4.5) were then evaluated for a total of 25 generations (default for GA used in this work). During the optimization parameters were allowed to vary within 5−6% of their initial values. The process was repeated, adjusting the parameter bounds, until the fitness (eq. 4.5) remained constant. There were numerous occasions in which false minima were identified along the parameter space, but these were eliminated on the basis of testing and evaluation of individual molecular errors present in the training set.

Upon evaluation of the parameters for H, C, and O it was discovered that only the carbohydrate systems yielded MUEs that were either comparable too or outperformed other SE methods, while systems possessing phosphates and amino acids possessed MUEs that were much larger than those of other SE methods. In order to correct for these errors a second set of parameterization was conducted in which parameters for H, C, and O were fixed to the values generated above and the only parameters optimized were those of N and P. Since C, H, and O were fixed we knew that the relaxation time for the carbohydrates would remain unchanged, as such the gas phase MD simulations were eliminated when optimizing the N and P parameters. In this case the optimum population size was found to be 512, mutation rate 0.50, elitism 0.05. This was run for a total of 25 generations (default for GA used in this work). Parameters for N and P were allowed to vary within 5−6% of their originals. Once again this process was repeated, adjusting the parameter bounds, until the smallest fitness was acquired. A final parameter optimization, in which H, C, N, O, and P parameters were optimized starting from the best set generated thus far, was conducted. With this set a population size of 100 (due to inclusion of MD

simulations), mutation rate 0.50, and elitism rate 0.05 was used for a total of 25 generations. Parameters were only allowed to vary within 0.5−1%. After tremendous evaluation the final set of parameters generated produced the newly established AM1/d-CB1 and AM1*-CB1 parameter sets.

The parameterization strategy described above differs from those of previously developed methods. Instead of using a specific reaction parameterization (SRP) strategy we parameterize with the aim of tackling a specific class of molecules (glycans) and the environment (amino acids and/or amino acid base pairs) within which they are known to exist in a chemical glycobiological landscape. We have named this process the *variable property optimization* (VPO) parameter approach.

4.6.1 Semi-empirical parameters

Table 4.2 provides a summary of parameters that are used, both in this work and in various SE Hamiltonians.

Table 4.2: Parameters used in various SE Hamiltonians

Parameter	Definition
U_{ss}, U_{pp}, U_{dd}	One-center one-electron integrals
ζ_{sn}, ζ_{pn}, ζ_{dn}	Internal orbital exponents
ζ_s, ζ_p, ζ_d	Slater orbital exponents
β_s, β_p, β_d	Two-center one-electron resonance integral
α_A	Core-core repulsion term
α_{AB}	Diatomic exponent core-core repulsion integral
x_{AB}	Diatomic core-core repulsion term
G_{ss}	s-s atomic orbital one-center two-electron repulsion integral
G_{sp}	s-p atomic orbital one-center two-electron repulsion integral
G_{pp}	p-p atomic orbital one-center two-electron repulsion integral
G_{p2}	p-p` atomic orbital one-center two-electron repulsion integral
H_{sp}	s-p atomic orbital one-center two-electron exchange integral
a_{nA}	Gaussian multiplier for the n^{th} Gaussian of atom A
b_{nA}	Gaussian exponent multiplier for the n^{th} Gaussian of atom A
c_{nA}	A radius of center of n^{th} Gaussian of atom A
F^0_{sd}, G^2_{sd}	Slater-Condon parameters

4.8 References

(1) Kayi, H. PhD Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.
(2) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(3) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, *9*, 408.
(4) Winget, P.; Clark, T. *J. Mol. Model.* **2005**, *11*, 439.
(5) Kayi, H.; Clark, T. *J. Mol. Model.* **2007**, *13*, 965.
(6) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 295.
(7) Kayi, H.; Clark, T. *J. Mol. Model.* **2009**, *15*, 1253.
(8) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 29.
(9) Kayi, H.; Clark, T. *J. Mol. Model.* **2010**, *16*, 1109.
(10) Kayi, H.; Clark, T. *J. Mol. Model.* **2011**, *17*, 2585.
(11) Nam, K.; Gao, J.; York, D. M. *J. Am. Chem. Soc.* **2008**, *130*, 4680.
(12) Wong, K.-Y.; Lee, T.-S.; York, D. M. *J. Chem. Theory Comput.* **2010**, *7*, 1.
(13) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(14) Leach, A. R. *Molecular Modelling: Principles and applications*; 2nd ed.; Prentice Hall, 2001.
(15) Linstrom, P.; Mallard, W. *NIST Chemistry WebBook, http://webbook.nist.gov/chemistry*; NIST Standard Reference Database Number 69: National Institute of Standards and Technology, Gaithersburg MD, 2003.
(16) Alexeev, Y.; Windus, T. L.; Zhan, C.-G.; Dixon, D. A. *Int. J. Quant. Chem.* **2005**, *102*, 775.
(17) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 2993.
(18) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243.
(19) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.
(20) Jalbout, A. F.; Adamowicz, L.; Ziurys, L. M. *Chem. Phys.* **2006**, *328*, 1.
(21) Giese, T. J.; Gregersen, B. A.; Liu, Y.; Nam, K.; Mayaan, E.; Moser, A.; Range, K.; Faza, O. N.; Lopez, C. S.; Lera, A. R. d.; Schaftenaar, G.; Lopez, X.; Lee, T.-S.; Karypis, G.; York, D. M. *J. Mol. Graphics Modell.* **2006**, *25*, 423.
(22) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173.
(23) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
(24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, Inc., Wallingford CT: 2009.
(25) Lewars, E. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*; 2nd ed.; Kluwer Academic Publishers, 2003.

(26) Almerindo, G. I.; Tondo, D. W.; Pliego, J. R. *J. Phys. Chem. A* **2003**, *108*, 166.
(27) Chandra, A. K.; Goursot, A. *J. Phys. Chem.* **1996**, *100*, 11596.
(28) Fu, Y.; Liu, L.; Li, R.-Q.; Liu, R.; Guo, Q.-X. *J. Amer. Chem. Soc.* **2003**, *126*, 814.
(29) Hudáky, P.; Perczel, A. *J. Phys. Chem. A* **2004**, *108*, 6195.
(30) Lopez, X.; Schaefer, M.; Dejaegere, A.; Karplus, M. *J. Amer. Chem. Soc.* **2002**, *124*, 5010.
(31) Magill, A. M.; Cavell, K. J.; Yates, B. F. *J. Amer. Chem. Soc.* **2004**, *126*, 8717.
(32) Moser, A.; Range, K.; York, D. M. *J. Phys. Chem. B* **2010**, *114*, 13911.
(33) Ozment, J. L.; Schmiedekamp, A. M. *Int. J. Quantum Chem.* **1992**, *43*, 783.
(34) Range, K.; López, C. S.; Moser, A.; York, D. M. *J. Phys. Chem. A* **2005**, *110*, 791.
(35) Range, K.; McGrath, M. J.; Lopez, X.; York, D. M. *J. Amer. Chem. Soc.* **2004**, *126*, 1654.
(36) Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.
(37) Hill, A. D.; Reilly, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 1031.
(38) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569.
(39) Naidoo, K. J.; Scientific Computing Research Unit, University of Cape Town: http://www.scientificomputing.com/.
(40) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(41) Thiel, W. *MNDO97*, version 5.0; University of Zurich, Zurich, Switzerland, 1998.
(42) Beck, B.; Horn, A.; Carpenter, J. E.; Clark, T. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1214.
(43) Groenhof, G. In *Biomolecular Simulations: Methods and Protocols (Methods in Molecular Biology)*; Monticelli, L., Salonen, E., Eds.; Springer Science and Business Media: New York, 2013; Vol. 924, p 43.
(44) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
(45) Palangsuntikul, R. PhD Thesis, Christian-Albrecht-University of Kiel, 2005.
(46) Lopez, X.; York, D. M. *Theor. Chem. Acc.* **2003**, *109*, 149.
(47) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231.

5. AM1/d-CB1: A semi-empirical method designed for QM/MM simulations of chemical glycobiology systems

The new parameters obtained for AM1/d-CB1 are presented and their accuracy is evaluated within the framework of the training set which was used during parameterization.

5.1 Results and Discussion

Adapting the parameterization strategy provided in Chapter 4 an optimum set of parameters for AM1/d-CB1 were acquired and these are listed in Table 5.1.

Table 5.1: Optimized AM1/d-CB1 parameters for Hydrogen, Carbon, Nitrogen, Oxygen and Phosphorus

Parameters	Hydrogen	Carbon	Nitrogen	Oxygen	Phosphorus
U_{ss}	-11.960909	-50.301531	-69.842739	-96.951432	-45.405057
U_{pp}		-38.793389	-55.880457	-77.905354	-41.533302
U_{dd}					-26.708704
ζ_s	1.052925	1.822969	2.351342	3.128851	2.058999
ζ_p		1.801099	2.033642	2.585827	2.214770
ζ_d					0.816679
β_s	-5.792945	-15.298988	-20.881617	-29.843262	-11.435826
β_p		-8.001275	-16.165663	-29.460458	-10.694210
β_d					-2.580718
α	3.026944	2.819744	3.185782	4.207192	2.050087
G_{ss}	13.808409	12.967197	11.847719	13.865036	14.263381
G_{pp}		11.063079	13.339745	14.481686	12.379996
G_{sp}		11.231690	12.735692	15.108816	5.769559
G_{p2}		9.872842	11.665818	12.449295	9.531268
H_{sp}		1.502380	4.683588	3.915720	1.272332
ζ_{sn}					2.069613
ζ_{pn}					1.485597
ζ_{dn}					1.139956
ρ_{core}					1.085029
G_{scale}	1.000000	1.000000	1.000000	1.000000	0.388294
FN_{11}	0.101830	0.075134	0.057001	0.228672	-0.334497
FN_{21}	5.891927	5.898126	4.339867	5.225437	3.202253
FN_{31}	1.175830	1.026976	1.283016	0.914621	1.020740
FN_{12}	0.065851	0.012140	0.023972	0.058956	-0.024098
FN_{22}	6.368911	6.956238	4.760398	7.537833	1.758030
FN_{32}	1.941724	1.664940	2.011604	1.516886	2.731363
FN_{13}	-0.034689	0.036407	-0.023463		-0.035212
FN_{23}	2.856686	6.263881	2.028720		4.902280
FN_{33}	1.625337	1.658710	1.961806		2.045419
FN_{14}		-0.002767			
FN_{24}		9.001121			
FN_{34}		2.817645			

The original parameters from which the AM1/d-CB1 parameters were derived are provided in Table A1 (Appendix A).

5.1.1 Key molecular properties to consider in chemical glycobiology

To accurately model the glycans in cellular systems the computation of specific properties important in chemical glycobiology must be used as metrics in the parameter optimization process. The following summarizes chemical features and molecular characteristics that we have taken into account during the parameterization process.

Ring Flexibility and Pucker. Reactions in chemical glycobiology involve the presence of either a 5- or 6-membered carbohydrate ring. Carbohydrate rings are conformationally flexible[1] although they are not as flexible as cycloalkanes e.g., cyclohexane.[2] It is this flexibility that leads to an exploration of ring pucker conformational space during the progression of hydrolysis, glycosylation and phosphorylation reactions. A ring that is too *stiff* or too *floppy* will not adapt to the important conformers needed in glycobiological reactions (see Scheme 1.1 and 1.2 of Chapter 1). We therefore monitor the effect of the new parameters on the 5- and 6-membered carbohydrate ring relaxation times. Further since ring puckering is a major driving force for chemical glycobiological reactions we place high priority on the proton affinities and electrostatic character (dipole moments) of the TS and other rings that are puckered away from the $^4C_1$ or $^1C_4$ chair conformers.

Bond polarization. During a hydrolysis or glycosylation reaction the carbohydrate is not only puckered away from its equilibrium (e.g., pyranose chair (C)) ring conformer but localized partial charges evolve on the oxygen and carbon atoms. This is the positively charged oxocarbenium ion. Moreover the oxocarbenium ion positions nucleophilic residues and leaving groups in the catalytic site. To improve the accuracy of modeling the oxocarbenium ion and the nucleophilic residues bond polarity has to be computed as closely to an *ab inito* result as possible. This is done by better calculating the molecular dipole moments and ionization potentials of pyranose half chairs (H), envelopes (E), boats (B), and skew (S) ring conformers as well as furanose twist (T) and envelop (E) ring conformers. To increase existing NDDO modeling of the nucleophilic interactions we follow more closely their molecular dipole moments and ionization potentials.

Amino acid contributions to glycan reactivity. Scheme 1.1 and 1.2 (Chapter 1) show that in glycosyltransferase and glycosidase, or glycosylase proton transfer plays an essential role in glycosyl transfer.[3,4] It is important to accurately model proton affinities of acid and basic groups involved in these reactions using a QM/MM method. The generic mechanism, in for example glycosyltransferases, following nucleophilic attack is to have one residue act as acid catalyst in promoting the departure of the leaving group while the other acts as a base catalyst to abstract a proton from the acceptor substrate. With this in mind we track the proton affinities of amino acid residues, common to catalytic domains, as an important property that AM1/CB1 is to model as accurately as possible.

5.1.2 Ring relaxation times

The ring relaxation time is a measure of the dynamic performance of the ring indicating the ease with which the carbohydrate may access a transition state conformation during the reaction (see Schemes 1.1-1.2, Chapter 1). Previously we had shown that the generalized NDDO methods aimed at organic systems (AM1 and PM3) produced carbohydrate rings with very low free energy barriers making the monosaccharide rings too flexible compared with Hartree-Fock level of theory or the SCC-DFTB method.[1,5] Furthermore the minimum pathway for the pyranose ring from C to H or E conformations mapped poorly against the SCC-DFTB method. The PM3CARB-1 method, that had been as well as those specifically parameterized for carbohydrates, performed equally poorly for furanose rings and marginally better for pyranose rings.

As mentioned in Chapter 4 (section 4.4.6) the carbohydrate ring relaxation times of the 5- and 6-membered sugar rings were included in the AM1/d-CB1 parameter optimization. The two sugar rings used for the parameterization were tetrahydrofuran (5-membered ring) and *β*-D-glucopyranose (6-membered ring). The relaxation times were obtained by fitting of the time correlation function (TCFs) using eq. 4.4 (Chapter 4). In the absence of long time *ab initio* simulations we used TCFs obtained from SCC-DFTB simulations as reference. The relaxation times for SCC-DFTB, the specially designed NDDO carbohydrate methods PM3CARB-1[6] and PM3^{MS},[7] AM1/d-PhoT[8] and AM1/d-CB1 are listed in Table 5.2 while AM1, PM3 and RM1 data are given in Table A2 (Appendix A).

Table 5.2: Ring relaxation times for molecules used in parameterization (picoseconds)

	SCC-DFTB[a]	PM3CARB-1	PM3^{MS}	AM1/d-PhoT	AM1/d-CB1
Tetrahydrofuran					
τ	0.47214	0.10223	0.10090	0.14734	0.22579
β-D-glucopyranose					
T	0.17384	0.17237	NONE	2.10800	0.14219

[a] Theoretical values obtained with gas-phase SCC-DFTB[5] MD simulations.[b] Correlation time could not be established since exponential fit was not possible with data generated from the dynamics run. τ corresponds to relaxation time for carbohydrate ring described in section 4.4.6 of Chapter 4. ^{NONE} implies that a relaxation time could not be obtained within the simulation time frame.

In addition to the ring relaxation times we calculated the root mean square deviations (RMSDs) of atoms that lie adjacent to the reference plane (Figure 1.6, Chapter 1). This gives an indication of the time it takes for a carbohydrate system to equilibrate. Average RMSDs for SCC-DFTB, PM3CARB-1, PM3^{MS}, AM1/d-PhoT and AM1/d-CB1 are provided in Figure 5.1. Individual atomistic RMSDs are shown in Figures A1-A2 (Appendix A).

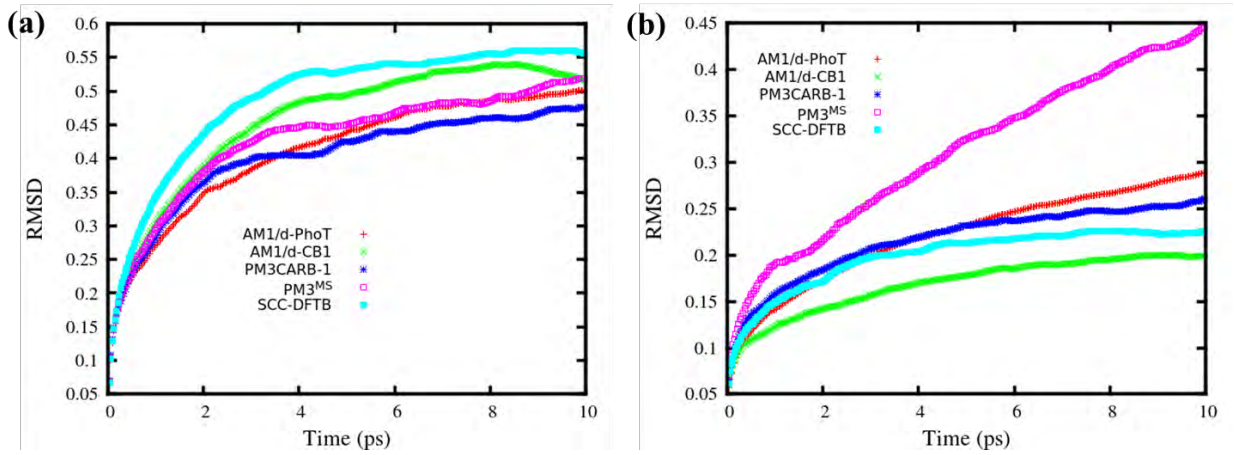


Figure 5.1: Average RMSD for (a) tetrahydrofuran and (b) *β*-D-glucopyranose.

For tetrahydrofuran all SE methods have much shorter relaxation times (~0.1 ps) than SCC-DFTB (Table 5.2) however, AM1/d-CB1's relaxation time (0.20484 ps) is closest to SCC-DFTB's ring relaxation time (0.47214 ps). However, the tetrahydrofuran AM1/d-CB1 model takes longer (6 ps) to equilibrate than the SCC-DFTB does (4.5 ps). The other methods do not reach equilibration within 10 ps. This is an indication that methods such as AM1/d-PhoT, PM3CARB-1 and PM3^{MS} have 5-membered rings that are too flexible.

For β-D-glucopyranose both AM1/d-CB1 and PM3CARB-1 relaxation times (0.17237 and 0.14219 ps, respectively) correlate perfectly with the dynamics of the SCC-DFTB ring model (0.17384 ps). However, AM1/d-CB1 equilibrates within 6.5 ps, which is longer than that of SCC-DFTB (4.5 ps) indicating a more flexible six membered ring. AM1/d-PhoT and $PM3^{MS}$ six membered ring dynamics correlates less well with SCC-DFTB. We were unable to compute a ring relaxation time for $PM3^{MS}$ since the data obtained from the simulation was too sporadic to fit. The RMSD plots (Figure 5.1b) confirm the poor behavior of both AM1/d-PhoT and $PM3^{MS}$ with a continually increasing average RMSD resulting in models that are not able to equilibrate within 10ps.

5.1.3 Gas phase proton affinities

As mentioned in Chapter 4 (section 4.4.4) determination of pK_a values is considerably challenging. In the current work we include both absolute (experimentally available data) and relative (DFT results) proton affinities (PA). The molecules used in the training set were grouped into molecular subclasses. A summary of these subclasses is provided in Table A3 of Appendix A. Table A4 (Appendix A) provides a comparison of calculated PAs and experimental data. For the reference proton affinities, we make use of the results from M06-2X calculations since it has been shown that this functional yields excellent PA results in comparison with experiments.[9-12]

As mentioned above acid and base residues in the catalytic domain of glycoenzymes are key to the success of glycan hydrolysis and glycosylation reactions. Therefore the parameters that produce accurate amino acid PAs is a priority in our *VPO* parameterization strategy. PAs of amino acids present in the training set correspond to the enthalpy change from a neutral amino acid species to an N-protonated species (+1 charge). Data for this protonation state was acquired from both high-level calculations[13] as well as experiment.[14] Single point SE calculations were done on the G3MP2 optimized geometries. AM1/d-CB1 and AM1/d-PhoT give the best performance for amino acids with MUEs of 4.3 and 4.4 kcal/mol, respectively (Table A4). Table A4 shows that for H_2O AM1/d-PhoT produces the smallest error of 4.7 kcal/mol, and $PM3^{MS}$ has an error of 7.2 kcal/mol, while all other SE methods have errors that are larger than 9 kcal/mol. For methanol it is AM1, PM3, and AM1/d-PhoT that have the smallest errors (2.2, -1.8, and 2.0 kcal/mol, respectively). Figure 5.2 illustrates the MUE of proton affinities (PAs) for the subclasses of molecules used during parameterization. The errors of AM1/d-CB1 are shown in

comparison for SE methods that have been used for carbohydrate modeling (PM3CARB-1, $PM3^{MS}$) or may be used in simulations of chemical glycobiology (AM1/d-CB1). Further since the AM1/d-PhoT Hamiltonian is foundational to the development of AM1/d-CB1 these results were of relevance in gauging the convergence toward the optimal AM1/d-CB1 parameter set. AM1/d-PhoT produces the smallest error for the organic molecules (Figure 5.2a) with an error of 3.3 kcal/mol (Table A4). In the case of phosphates AM1/d-PhoT and AM1/d-CB1 give the smallest errors (9.0 and 7.5 kcal/mol, respectively) underscoring the importance of including *d*-orbitals on phosphorus. The molecule contributing the largest error for the phosphates is $(OCH_3)_2(OH)PO$ with an error of 42.5 and 30.4 kcal/mol for AM1/d-PhoT and AM1/d-CB1, respectively.

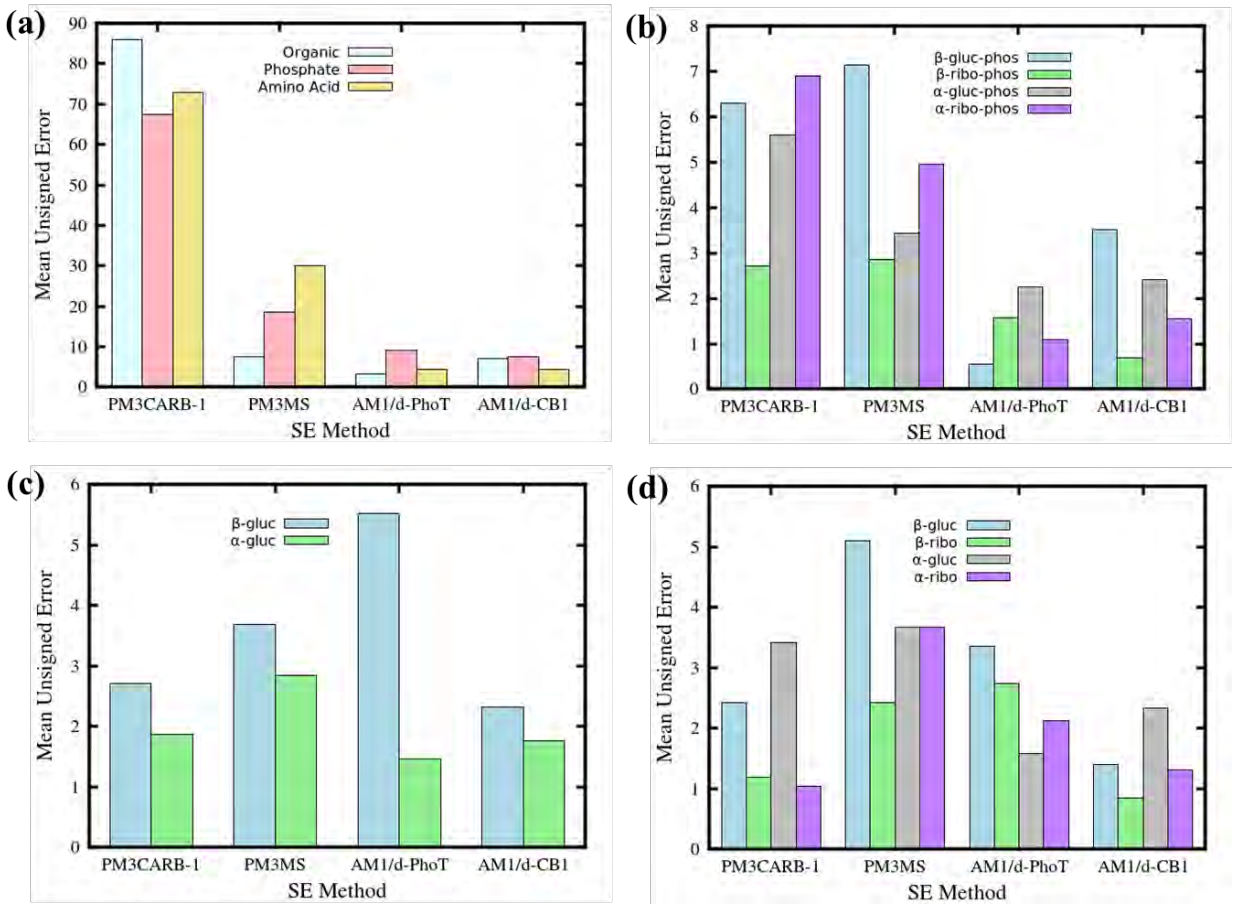


Figure 5.2: Mean unsigned errors for gas phase proton affinities of (a) organic molecules, phosphates and amino acids, (b) carbohydrate chair phosphorylated conformers, (c) carbohydrate chair conformers of glucopyranose and (d) puckered carbohydrate transition state conformers of glucopyranose and ribofuranose.

AM1/d-PhoT and AM1/d-CB1 yield the smallest errors for carbohydrate-phosphate PAs (Figure 5.2b). A more detailed look at the individual systems reveals that the lowest MUE for the β-D-glucopyranose-phosphate and α-D-ribofuranose-phosphates is given by AM1/d-PhoT with values of 0.54 and 1.09 kcal/mol, respectively (Table A5). The MUE of α-D-glucopyranose-phosphate is similar for AM1/d-PhoT and AM1/d-CB1 with values of 2.26 and 2.41 kcal/mol, respectively. Methods that have previously been parameterized for carbohydrates (PM3 and PM3CARB-1) appear less suitable for modeling of a proton acceptance by a phosphate, which is an important facet for chemical glycobiological reactions (Chapter 1, Scheme 1.1).

The protonation of carbohydrates is important in glycan reactions at the anomeric carbon (for example reactions shown in Scheme 1.1 of Chapter 1). PAs for minimum energy conformers ($^4C_1$ and $^1C_4$) of glucopyranose for AM1/d-CB1 (Figure 5.2c) surpasses all other methods. For the α- anomers the method gives an error (1.76 kcal/mol) that is slightly higher than that of AM1/d-PhoT (1.47 kcal/mol).

Computing the PAs of non-chair conformers that are often present in transition state configurations is a priority molecular property in this *VPO* strategy. To maintain the structural integrity of the carbohydrate conformers only single point calculations were performed.[15] AM1/d-CB1 models of β-D-glucopyranose and β-D-ribofuranose give PAs of 1.41 kcal/mol and 0.85 kcal/mol, respectively that are better than other NDDO methods (Table A6). In the remaining cases AM1/d-CB1 is second only to AM1/d-PhoT (1.59 kcal/mol) and PM3CARB-1 (1.05 kcal/mol) for α-D-glucopyranose (2.34 kcal/mol) and α-D-ribofuranose (1.31 kcal/mol) conformers, respectively (Table A7).

5.1.4 Dipole moments

During the parameterization of AM1/d-CB1, the DFT dipole moments were used as reference data. The molecular electrostatic potential correlates strongly with dipole moment as a result we placed high value to the accuracy of this property (see Chapter 4, Table 4.1). Therefore overall for each molecular subclass the errors for AM1/d-CB1 are small although not the smallest in every case.

The amino acid residue dipole moments were a priority property in the *VPO* strategy as an accurate representation of the electrostatic potential within the catalytic domain of the glycoenzymes as well as within the protein binding site of carbohydrate binding proteins are

essential. Therefore, AM1/d-CB1 gives the best dipole moments for amino acids as well as organic molecules (Figure 5.3a) with MUEs of 0.26 and 0.24 debye, respectively (Table A8). $PM3^{MS}$ performance (MUE of 0.38 debye) is closest to AM1/d-CB1 (Table A8) for the organics. AM1/d-PhoT and PM3CARB-1 gives larger errors for organics with MUEs of 0.55 and 0.56 debye, respectively. In the case of the amino acids it is AM1/d-PhoT that yields an error close to AM1/d-CB1 (0.33 debye). The methods parameterized for carbohydrates (PM3CARB-1 and $PM3^{MS}$) by comparison perform less well for amino acids giving MUEs of 0.74 and 0.58 debye, respectively (Table A8).

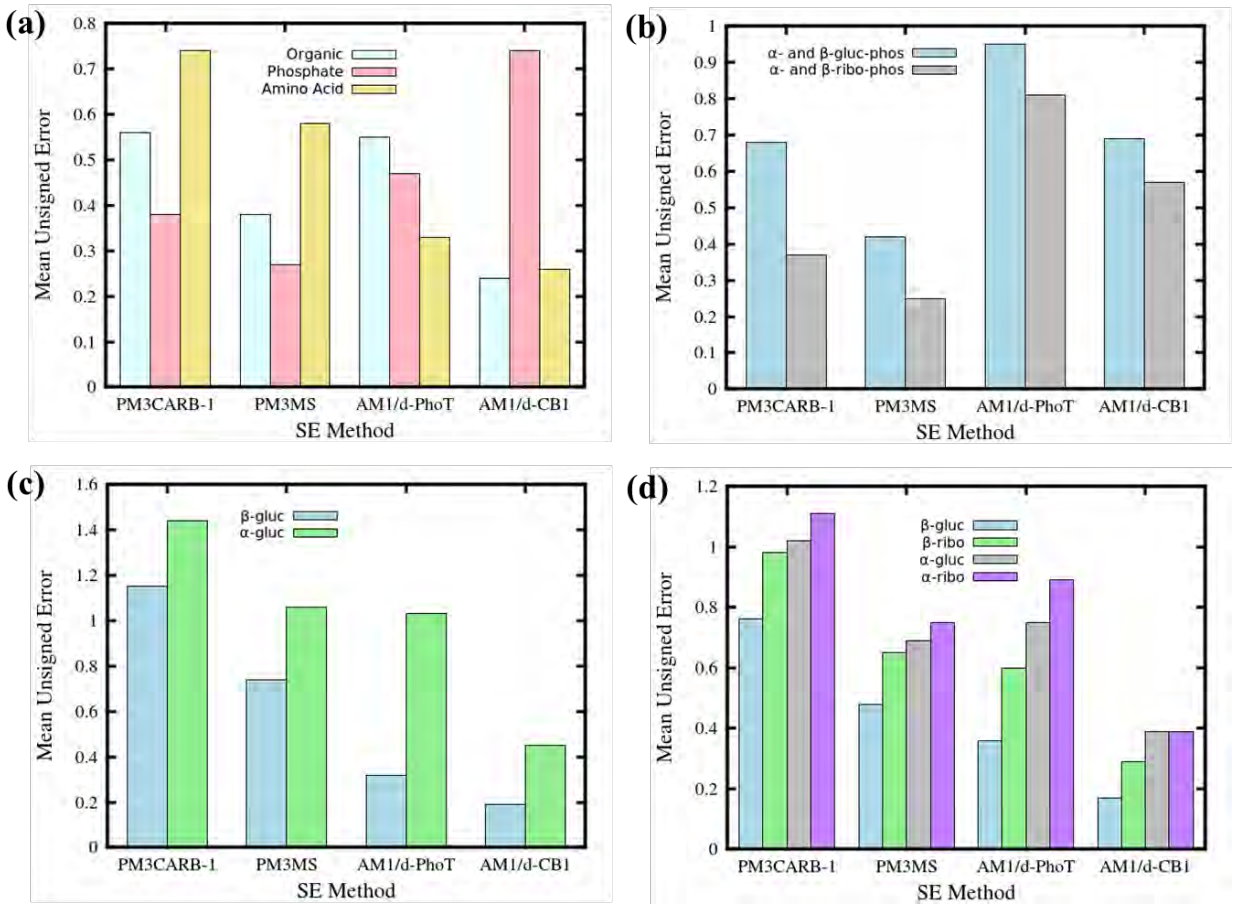


Figure 5.3: Mean unsigned errors of dipole moments for (a) organic molecules, phosphates and amino acids, (b) carbohydrate chair phosphorylated conformers, (c) carbohydrate chair conformers of glucopyranose and (d) none equilibrium puckered carbohydrate ring conformers of glucopyranose and ribofuranose.

Interestingly, the PM3^{MS} and PM3CARB-1 methods that do not incorporate *d*-orbital character into phosphorus, give smaller dipole moment errors for the phosphate molecules than do the AM1/d-PhoT and AM1/d-CB1 methods that do include *d*-orbital character into phosphorus. While the error for AM1/d-CB1 is not the lowest for phosphates it is relatively small (MUE of 0.74 debye). Removal of two species that are not significant in chemical glycobiology ($P(CH_3)_3$ and $(CH_3)_3PO$) produces a much smaller MUE for AM1/d-CB1 (0.15 debye).

The MUEs for the carbohydrate-phosphate systems are <1 debye for all four methods shown in Figure 5.3b. While PM3CARB-1 and PM3^{MS} produce errors that are smaller, these methods apply only an *sp* basis onto hypervalent atoms such as phosphorus making the electronic model inaccurate. AM1/d-CB1 performs better than AM1/d-PhoT for example the MUEs for glucopyranose-phosphate are 0.69 and 0.95 debye, respectively, and 0.57 and 0.81 debye, respectively, (Table A8 in Appendix A) for the ribofuranose-phosphate.

Of greater importance to this work is the modeling of carbohydrate conformations that are commonly found in transition state structures or along the reaction coordinate. We computed the dipole moments for *C*, *H*, *E, B* and *S* molecular *β*-D-glucopyranose ring conformers (Tables A9-A10 in Appendix A) and show the MUEs (Figures 5.3c-d). The carbohydrate *H*, *E, B* and *S* rings were a particular priority of the *VPO* strategy, therefore, AM1/d-CB1 shows the smallest deviation from the M06-2X/6-311++G(3df,2p) computed dipoles. For the minimum energy chair α- and *β*-D-glucopyranose conformers AM1/d-CB1 produces the smallest MUE of 0.45 and 0.19 debye, respectively. A similar trend is found for the transition state glucopyranose conformers with AM1/d-CB1 producing MUEs of 0.39 and 0.17 debye, respectively. The errors for the dipole moments were reduced during the final parameter refinement; however, this reduction did cause an increase in the errors for the ionization potential, as shall be noted in the section that follows.

5.1.5 Ionization potential

As mentioned in Chapter 4 (Section 4.4.3) there are two types of ionization potential and for the purposes of this work the *vertical* ionization potential (IP) was used. All cationic forms of species listed in Tables A11-A13 were computed as single point calculations using M06-2X/6-311++G(3df,2p) optimized neutral molecule geometries.

AM1/d-PhoT and AM1/d-CB1 give the smallest errors for all molecular subclasses (Figure 5.4) with PM3CARB-1 and $PM3^{MS}$ giving considerably larger MUEs. In general the AM1/d-CB1 errors are second only to AM1/d-PhoT. Ionization potentials play an important role in the modeling of nucleophilic reactions. We therefore prioritized the ionization potentials of amino acids in the *VPO* strategy. The result is that AM1/d-CB1 outperforms all the methods for the chemically glycobiologically significant amino acids used in the training set presented in Figure 5.4a with an error of 0.20 eV (Table A11).

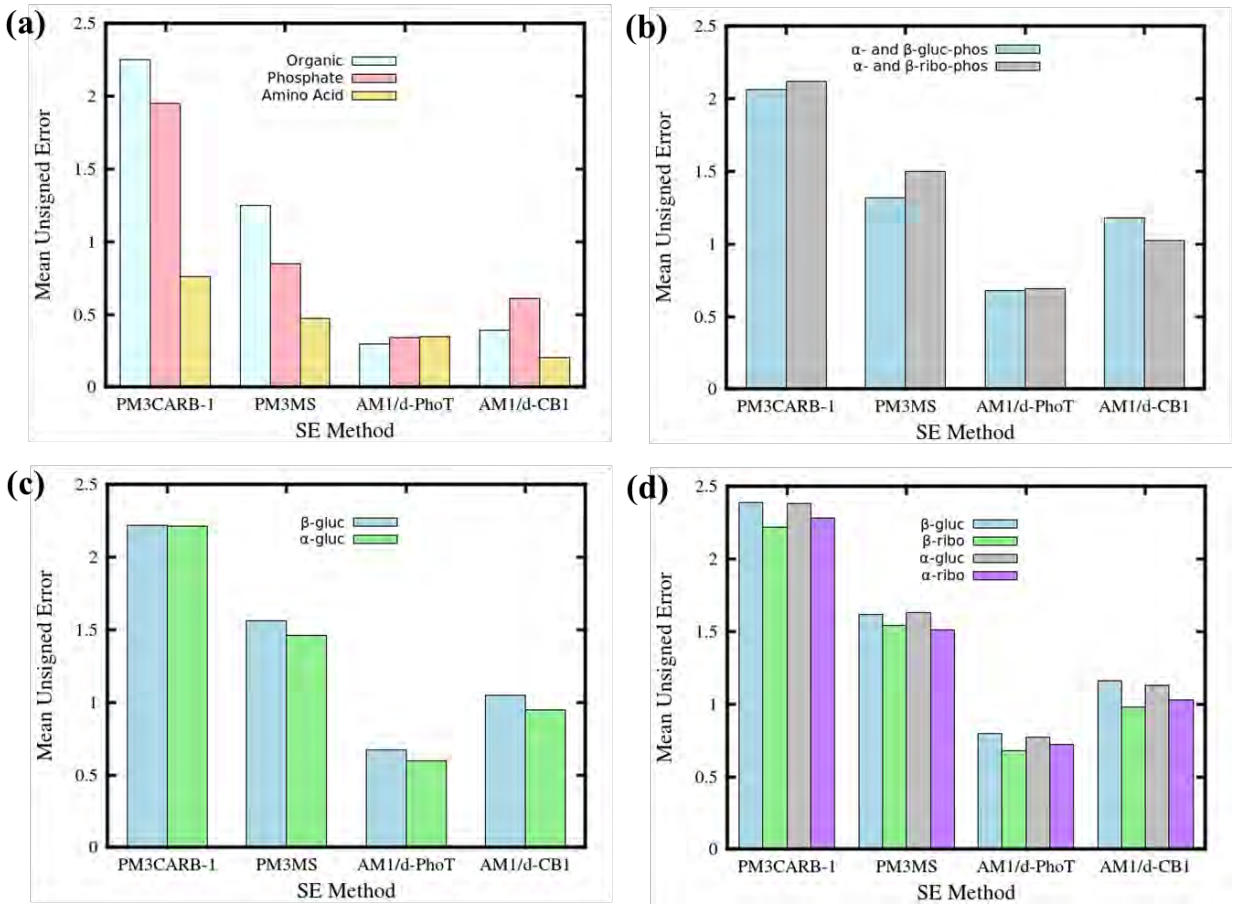


Figure 5.4: Mean unsigned errors of ionization potentials for (a) organic molecules, phosphates and amino acids, (b) carbohydrate chair phosphorylated conformers, (c) carbohydrate chair conformers of glucopyranose and (d) none equilibrium puckered carbohydrate conformers of glucopyranose and ribofuranose.

Unlike with properties mentioned above, AM1/d-CB1 produces errors that are second to AM1/d-PhoT for the minima and transition state conformers of carbohydrate rings (Figures 5c-d). The larger errors stem from both an improvement in the dipole moments during parameter refinement, and the smaller weights used for the ionization potential during parameterization. A higher weighting for the ionization potential has been tested, but resulted in an increases in the errors of other properties considered during this work.

5.1.6 Interaction energies

The DFT energies for bimolecular complexes were obtained using M06-2X/6-31+G(df). The purpose of this work is to develop a method, which will accurately model reactions important in glycobiology, and such reactions would involve hydrogen bonding with the surrounding water environment. A number of hydrogen bonded dimers were used in the parameterization of AM1/d-CB1. Results for the various hydrogen bond dimers are provided in Table 5.3 and Table A14 (Appendix A).

Table 5.3: Experimental and Theoretical interaction energies for molecules used in parameterization (kcal/mol)

Molecule	Reference		Error			
	Exp	DFT[b]	PM3CARB-1	$PM3^{MS}$	AM1/d-PhoT	AM1/d-CB1
H_2O:H_2O	5.00[a]	-5.18	1.82	0.39	0.89	2.91
H_2O:CH_3OH		-5.17	1.96	0.67	1.83	2.72
H_2O:PO_3^-		-15.90	5.62	7.52	0.46	-0.77
H_2O:$H_2PO_4^-$		-18.20	6.75	8.83	-0.19	-2.41
H_2O:HPO_4^{2-}		-33.27	3.50	7.26	3.56	-1.62
MUE (vs DFT)			**3.93**	**4.93**	**1.38**	**2.08**
MSE (vs DFT)			**3.93**	**4.93**	**1.31**	**0.17**

[a] Experimental value obtained from Feyereisen et al.[16] [b] The DFT interaction energies where computed with M06-2X/6-31+G(df). All errors are computed as $\Delta H_{\text{int}}^{calc} - \Delta H_{\text{int}}^{ref}$.

AM1/d-PhoT gives the smallest MUE of 1.38 kcal/mol followed by AM1/d-CB1 where the MUE (2.08 kcal/mol) is approximately half that of PM3CARB-1 and $PM3^{MS}$. A closer look at the individual errors shows that $PM3^{MS}$ yields the smallest errors for the water dimer and water-methanol interaction, while giving the highest MUEs for water phosphate complexes. In the case of water-phosphate interactions AM1/d-CB1 compares substantially better with the DFT calculations than does PM3CARB-1 or $PM3^{MS}$. Although some of the errors acquired with

AM1/d-CB1 may appear large it should be noted that the results given in Table 5.3 are being compared to DFT and M06-2X functional that implicitly includes corrections for dispersion and hydrogen bonding. Interaction energies for the complexes shown in Table 5.3 computed using AM1, PM3 and RM1 are tabulated in Table A14 of Appendix A.

5.1.7 Heats of formation (ΔH_f)

In the optimization process, experimental heats of formation were used as the target where available, whereas DFT results from M06-2X calculations were used when experimental results were absent. In our optimization of AM1/d-CB1 parameters we assigned the lowest weighting to the heat of formation property. We compare carbohydrate specific methods (PM3CARB-1, $PM3^{MS}$ and AM1/d-CB1) and AM1/d-PhoT that includes a *d*-orbital treatment of phosphorus. The comparative performance of these methods for general organic, phosphates and amino acid molecular subclasses often present in chemical glycobiology events are shown in Figure 5.5a. Here AM1/d-PhoT and AM1/d-CB1 produce the smallest errors for organic systems with AM1/d-PhoT yielding a MUE of 7.4 kcal/mol, while AM1/d-CB1 has an error of 10.7 kcal/mol. When introducing longer aliphatic chains into the alcohols i.e, ethanol (C_2H_5OH), propanol (C_3H_7OH) and finally butanol (C_4H_9OH), AM1/d-CB1's heat of formation performance declines (8.5, 15.6 and 22.8 kcal/mol, respectively). This is a result of the low weighting attached to ΔH_f during the optimization. It should be noted however, that a similar trend is observed for AM1/d-PhoT with errors of 6.2, 12.0 and 18.0 kcal/mol, respectively. AM1/d-CB1 errors for the phosphate molecules[17] (Figure 5.5a) are very similar to those of AM1/d-PhoT. The error produced for the amino acids with AM1/d-CB1 is 15.0 kcal/mol (Table A15 in Appendix A), which is two orders in magnitude lower than that of AM1/d-PhoT (30.6 kcal/mol).

AM1/d-CB1 has mixed success for carbohydrate-phosphate systems (Figure 5.5b). The AM1/d-CB1 heats of formation results for the β- anomer of glucopyranose are modelled better by AM1/d-CB1 (MUE 4.33 kcal/mol) than AM1/d-PhoT (MUE 6.78). Although $PM3^{MS}$ and PM3CARB-1 produce better results than AM1/d-CB1 with errors of 3.64 and 3.77 kcal/mol, respectively, it is important to note that these methods do not incorporate *d*-orbitals onto phosphorus making these methods less accurate for the modeling of hypervalent species. For β-D-ribofuranose AM1/d-CB1 surpasses the other methods by 3 to 4 orders in magnitude, with a MUE of 1.44 kcal/mol (Table A16). For the α- anomer of glucopyranose it is AM1/d-PhoT that

produces the lower errors of 1.48 kcal/mol, while the α- anomer of ribofuranose is modelled similarly for both AM1/d-PhoT and AM1/d-CB1 with MUE of 1.07 and 1.18 kcal/mol.

For the equilibrated chair conformations of glucopyranose (Figure 5.5c) AM1/d-CB1 produces the most accurate results with errors of 0.2 and 0.04 kcal/mol for the β- and α-anomers, respectively (Tables A17-A18, Appendix A). The MUEs for heats of formation of carbohydrate conformers puckered away from the equilibrated chair conformations are shown in Figure 5.5d. In β- anomers of both glucopyranose and ribofuranose systems AM1/d-CB1 outperforms all other methods in computing the heats of formation of possible transition state puckered rings. A minor deficiency is its performance that lags behind that of PM3CARB-1 and $PM3^{MS}$ for the α-ribofuranose.

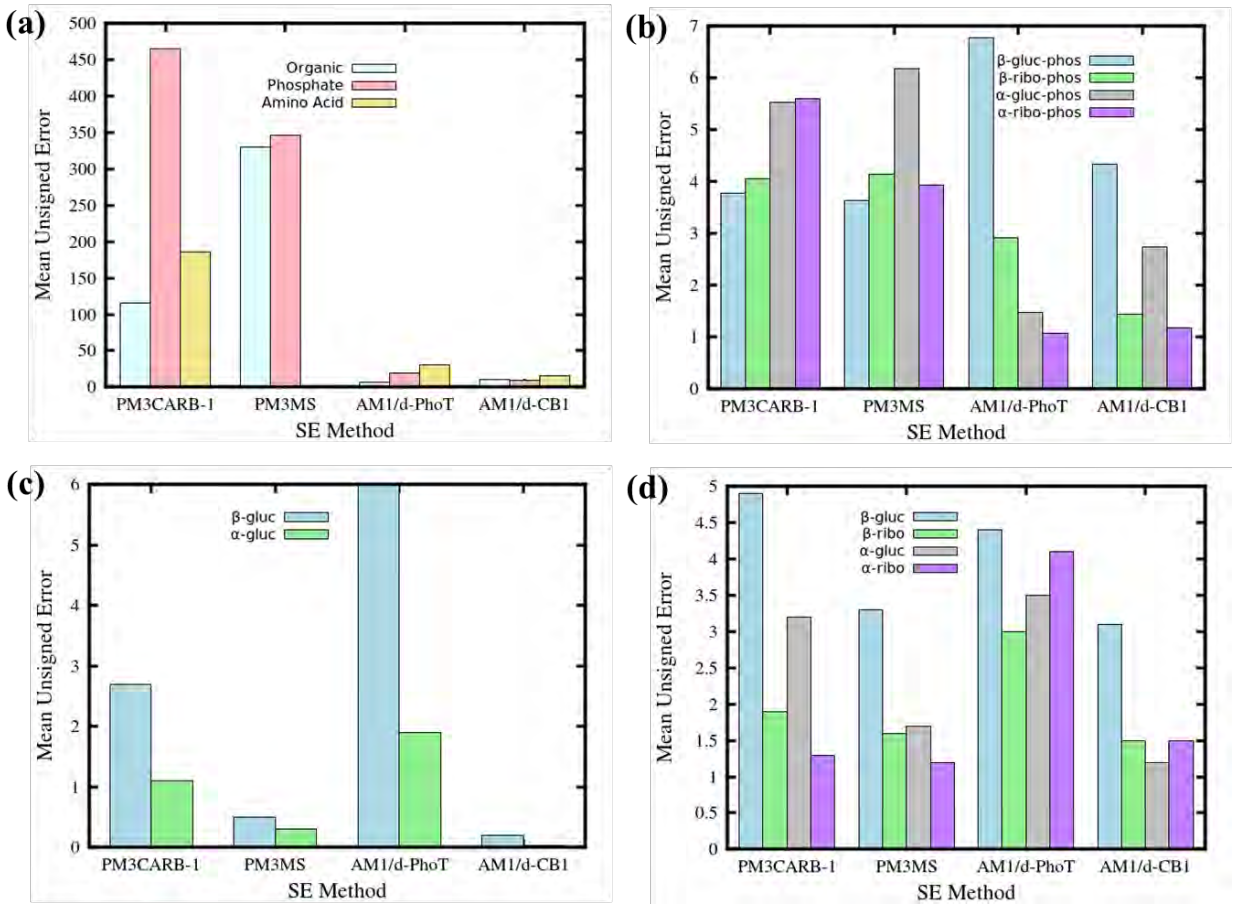


Figure 5.5: Mean unsigned errors for heats of formation for (a) organic molecules, phosphates and amino acids, (b) carbohydrate chair phosphorylated conformers, (c) carbohydrate chair conformers of glucopyranose and (d) none equilibrium puckered carbohydrate conformers of glucopyranose and ribofuranose.

5.2 Conclusion

A parameterization of the AM1/d Hamiltonian initiated from RM1 and AM1/d-PhoT models has been conducted with the aid of a genetic algorithm. The H, C, N, O, and P atoms were tuned using a *variable property optimization* parameter approach that prioritizes selective molecular classes for each property (dipole moment, heat of formation etc.) in addition to weighting properties differently to achieve the goal deriving a parameter set that is capable of modeling a glycan as well as its immediate environment in a chemical glycobiological context. We called this model AM1/d-CB1. In optimizing the semi-empirical parameters for these atoms we prioritized the dynamic performance of ring pucker and key elements of the glycoenzymes reaction class such as proton affinity (commonly associated with acid/base catalysis) and ionization energies (commonly associated with formation of oxocarbenium ions).

Computing accurate transition state properties for glycans in enzymatic reactions has been a universal failing of many SE methods. The property prioritization of ring puckering dynamics, the dipole moments and heats of formation of non-equilibrium ring conformers as well as the proton affinities, heats of formation and dipole moments of amino acids was central to the development of AM1/d-CB1. Proton affinities, dipole moments, ionization potential as well as heats of formation for transition state ring carbohydrate puckered conformations revealed that AM1/d-CB1 performs better than other SE methods that may be used for simulating glycoenzymes (glycosyltransferase, glycosidase, and glycosylase) catalyzed chemical reactions. However, the AM1/d model suffers from historically poor NDDO descriptions of hydrogen bond and dispersion interactions. Corrections to these deficiencies are currently under development for AM1/d-CB1 that will further improve its description of glycan reactivity as studied in chemical glycobiology.

The performance of AM1/d-CB1 across the many role players in chemical glycobiological reactions is presented in Chapter 6. There an evaluation of carbohydrate free energy pucker surfaces and volumes, phosphate reactions, and base pair associations is reported.

5.3 References

(1) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.
(2) Biarnés, X.; Ardèvol, A.; Iglesias-Fernández, J.; Planas, A.; Rovira, C. *J. Amer. Chem. Soc.* **2011**, *133*, 20301.
(3) Kapitonov, D.; Yu, R. K. *Glycobiology* **1999**, *9*, 961.

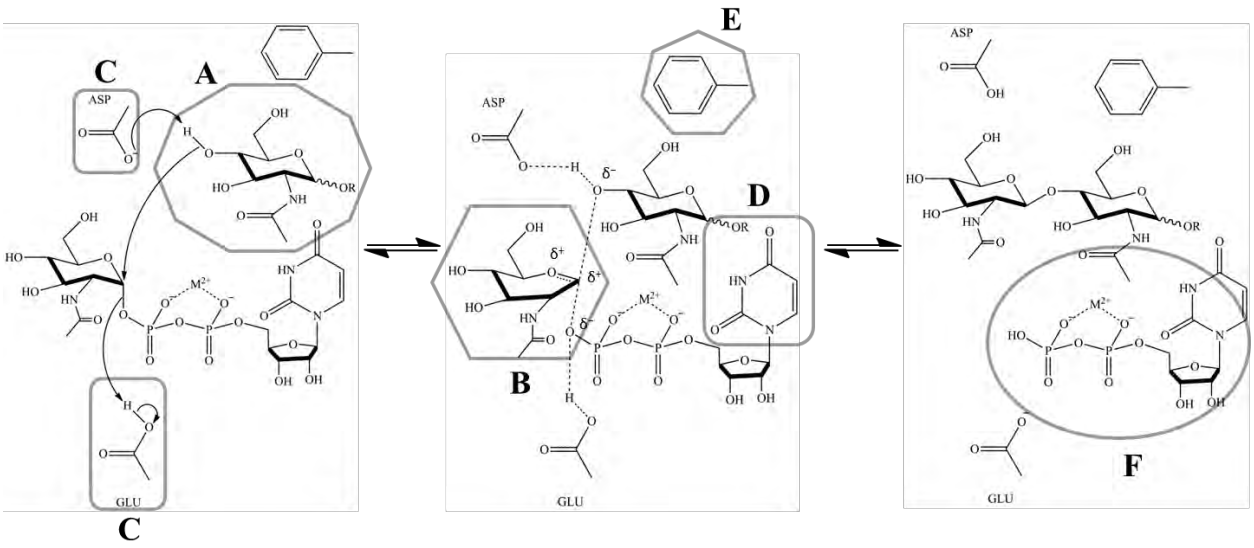
(4) Rye, C. S.; Withers, S. G. *Curr. Opin. Chem. Biol.* **2000**, *4*, 573.
(5) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. J. *J. Phys. Chem. B* **2001**, *105*, 569.
(6) McNamara, J. P.; Muslim, A.-M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, *394*, 429.
(7) Mane, J. Y.; Klobukowski, M. *Chem. Phys. Lett.* **2010**, *500*, 140.
(8) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(9) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
(10) Walker, M.; Harvey, A. J. A.; Sen, A.; Dessent, C. E. H. *J. Phys. Chem. A* **2013**, *117*, 12590.
(11) Peverati, R.; Truhlar, D. G. *Phil. Trans. R. Soc. A* **2014**, *372*.
(12) Brás, N. F.; Perez, M. A. S.; Fernandes, P. A.; Silva, P. J.; Ramos, M. J. *J. Chem. Theory Comput.* **2011**, *7*, 3898.
(13) Gronert, S.; Simpson, D. C.; Conner, K. M. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2116.
(14) Linstrom, P.; Mallard, W. *NIST Chemistry WebBook, http://webbook.nist.gov/chemistry;* NIST Standard Reference Database Number 69: National Institute of Standards and Technology, Gaithersburg MD, 2003.
(15) Jalbout, A. F.; Adamowicz, L.; Ziurys, L. M. *Chem. Phys.* **2006**, *328*, 1.
(16) Feyereisen, M. W.; Feller, D.; Dixon, D. A. *J. Phys. Chem.* **1996**, *100*, 2993.
(17) Alexeev, Y.; Windus, T. L.; Zhan, C.-G.; Dixon, D. A. *Int. J. Quant. Chem.* **2005**, *102*, 775.

6. The performance of AM1/d-CB1 for Chemical Glycobiology QM/MM simulations: Evaluating Carbohydrate ring pucker, phosphate reactions, amino acid binding and base pair associations

Testing of the AM1/d-CB1 parameters presented in Chapter 5 is conducted to evaluate the methods accuracy in modelling systems that are of a chemical glycobiological nature. This chapter is based on work provisionally accepted for publication in the Journal of Chemical Theory and Computation.

6.1 Results and Discussion

The accuracy of computer calculations relies on accurate models. When modeling reactions in chemical glycobiology a semi-empirical (SE) method must accurately model; A) the molecular structure of the monosaccharide, B) conformational (particularly ring puckering) and electronic transition (formation of oxocarbenium ion) of monosaccharides C) the ability of amino acids to accept and donate protons, D) the interactions involved when a base hydrogen bonds and/or π-stacks with molecules in the catalytic domain or with other bases, E) the interactions of a sugar with neighboring aromatic rings and F) the barrier heights required in order for a phosphate group to leave resulting in extended carbohydrate chain formation (oligosaccharides). In Chapter 5 we focused on the development of the semi-empirical (SE) AM1/d/CB1 method to specifically model glycans and more generally biochemical processes of interests in chemical glycobiology taking the above mentioned features into account (Scheme 6.1).[1] Here we evaluate the performance of AM1/d-CB1 and draw comparisons to NDDO methods (AM1,[2] PM3,[3,4] PM3CARB-1,[5] PM3^{MS} [6] and RM1[7]). Since most of these methods are incapable of correctly modeling hypervalent atoms such as phosphorus we include AM1/d-PhoT,[8] SCC-DFTB[9] and M06-2X[10] to further measure the relative performance of AM1/d-CB1



Scheme 6.1: Mechanism for inverting glycosyltransferase reaction involving a UDP-GlcNAc donor and GlcNAc-containing acceptor substrates. Labels represent: A – carbohydrate structure, B – carbohydrate ring pucker, C – proton accepting and donating amino acids, D – base pair interactions, E – carbohydrate-aromatic π stacking and F – phosphate leaving group.

6.1.1 Carbohydrate structure

There are nine monosaccharides that form the basic alphabet upon which the mammalian glycome (all the carbohydrates in an organism) is constructed. We used the M06-2X/6-311++G(3df,2p) (DFT) level of theory to get optimized structures for these nine monosaccharides (Figure B1, Appendix B). The relative heats of formation acquired from DFT, where all energies were computed relative to *β*-D-glucose, are compared to those generated from the SE calculations (Tables B1-B2, Appendix B). The structures obtained from DFT were used in both single point and geometry optimization calculations from which heats of formation were calculated. Methods that have been specifically parameterized for carbohydrates (PM3CARB-1 and $PM3^{MS}$) give rise to the largest Mean Unsigned Errors (MUEs), with single point errors of 104.29 and 194.45 kcal/mol, respectively (Table B1). The MUEs, 104.79 and 196.31 kcal/mol respectively, do not improve following geometry optimization (Table B2). AM1/d-CB1 substantially outperforms all of the methods (Figure 6.1) exhibiting errors of 3.64 and 5.02 kcal/mol for single point and geometry optimized structures, respectively (Tables B1-B2).

A comparison of the optimized coordinates with those obtained from the DFT simulations were used to compute the root mean square deviations (RMSD) of the individual monosaccharides (Table B3). AM1, PM3, RM1, and AM1/d-CB1 minimize to geometries that are very similar to those of DFT. Interestingly the methods parameterized for carbohydrates (PM3CARB-1 and $PM3^{MS}$) give the poorest optimized structures.

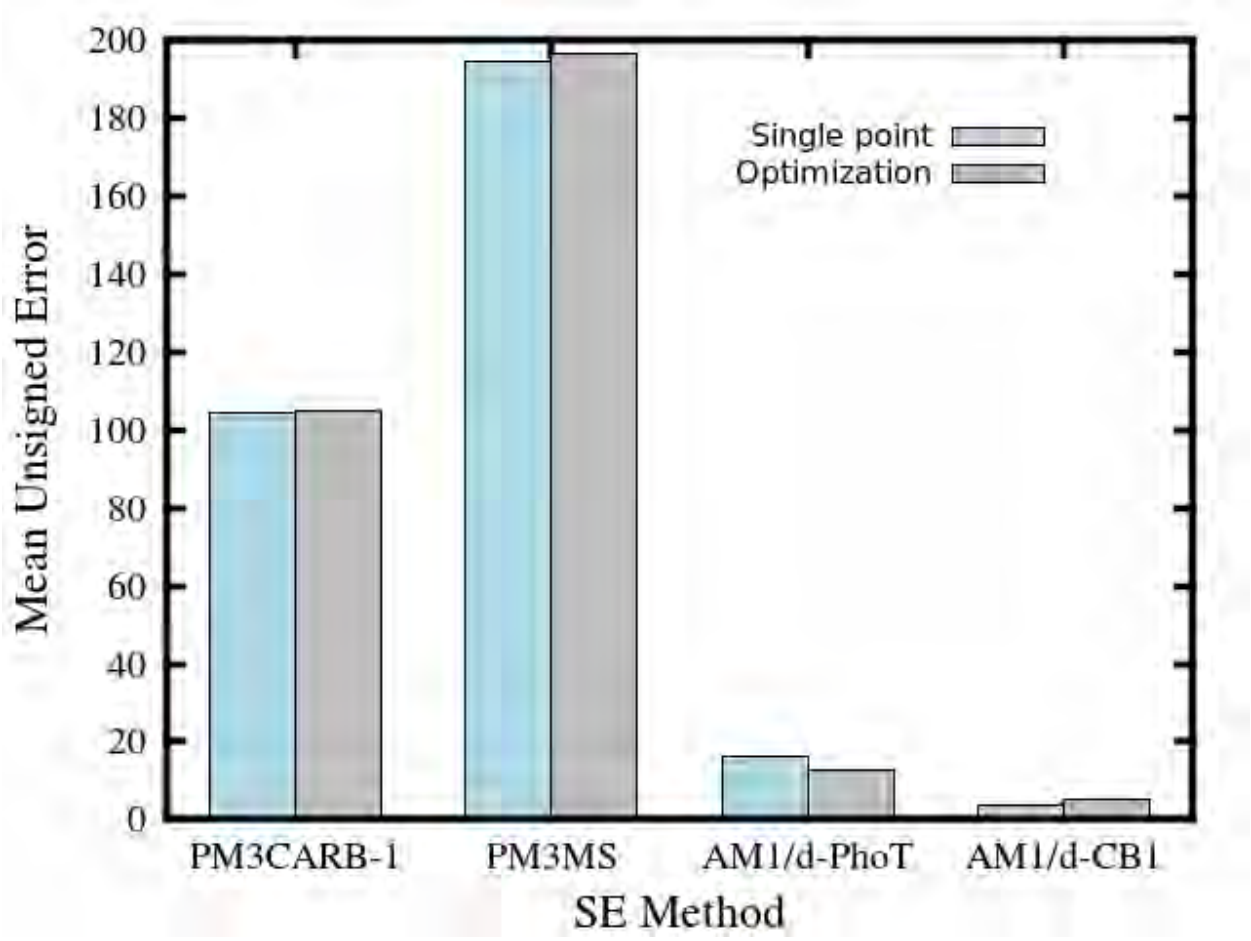


Figure 6.1: Relative mean unsigned errors in heats of formation (kcal/mol) for nine monosaccharaide using both single point and geometry optimization, on the DFT optimized structures.

6.1.2 Carbohydrate ring pucker from free energy simulations

The generalized free energy approach termed Free Energy from Adaptive Reaction Coordinate Forces (*FEARCF*)[11-13] was used with various SE based methods (including AM1/d-

CB1) to calculate the free energy of ring pucker[12] surface for β-D-ribofuranose and free energy ring pucker volume for β-D-glucopyranose. The accuracy of the computed reaction mechanism depends on the ability of the model to simulate ring pucker (Scheme 6.1, label B). The simulations conducted in this work followed the same methodology as described earlier[14] with the only differences being that CHARMM[15] v35b5 was used, instead of v33b2, and the QM MD simulations were conducted with the CHARMM/MNDO97[16] interface.

6.1.2.1 Ribofuranose

To compute the furanose ring pucker conformations the ring is subdivided into a reference plane and two rotatable planes using a triangular tessellation method.[17,18] The free energy of pucker is then computed as a function of the angles (θ_0, θ_1) that the rotatable planes make with the reference plane.[12,14,19] Canonical conformers of the furanose ring i.e., envelopes (E) and twists (T) are illustrated in terms of nodes (θ_0, θ_1) in Figure 6.2(a) significantly distanced away from the center of the surface (0,0) that corresponds to a flattened ring.

The Hartree-Fock (HF) free energy of pucker surface had been computed using 6-31G basis set (Figure 6.2b).[14] HF free energy computations require vast amounts of compute cycles as a result the surface is not converged. Nonetheless, the HF free energy of pucker for the furanose ring does reveal distinct minima interpretable as canonical puckered conformers. The global minimum (-35^o, 30^o) is a $^3T_2/^3E$ conformer. A second minimum exists at (27.5^o, -7.5^o) that is approximately 0.95 kcal/mol higher in energy than the global minimum.

Previously we compared AM1, PM3, PM3CARB-1 and SCC-DFTB free energy of pucker surfaces.[14] We showed that common to all NDDO methods is a large minimum energy well with no distinct global minimum. More seriously is the lack of energetic differentiation on the free energy surface (FES) that is evidence of discrimination between different puckering conformers. The shapes of NDDO (AM1, PM3, or PM3CARB-1) pucker FES' are indicative of furanose ring models that are flexible, can pucker relatively easily at room temperature and conformationally indiscriminate. For example $PM3^{MS}$ (Figure 6.2d) has a global energy minimum of (-2.5^o, 17.5^o) that is closes to the planar conformer with the nearest canonical conformer being ^1E. As with other SE methods $PM3^{MS}$ bowl shaped FES can access several conformers where the canonical ^1T_o, E_o, ^4T_o, ^4E lie on the periphery within 0.5 kcal/mol of the near flattened favoured ring conformer. This continues to be the case for AM1/d-PhoT (Figure 6.2e).

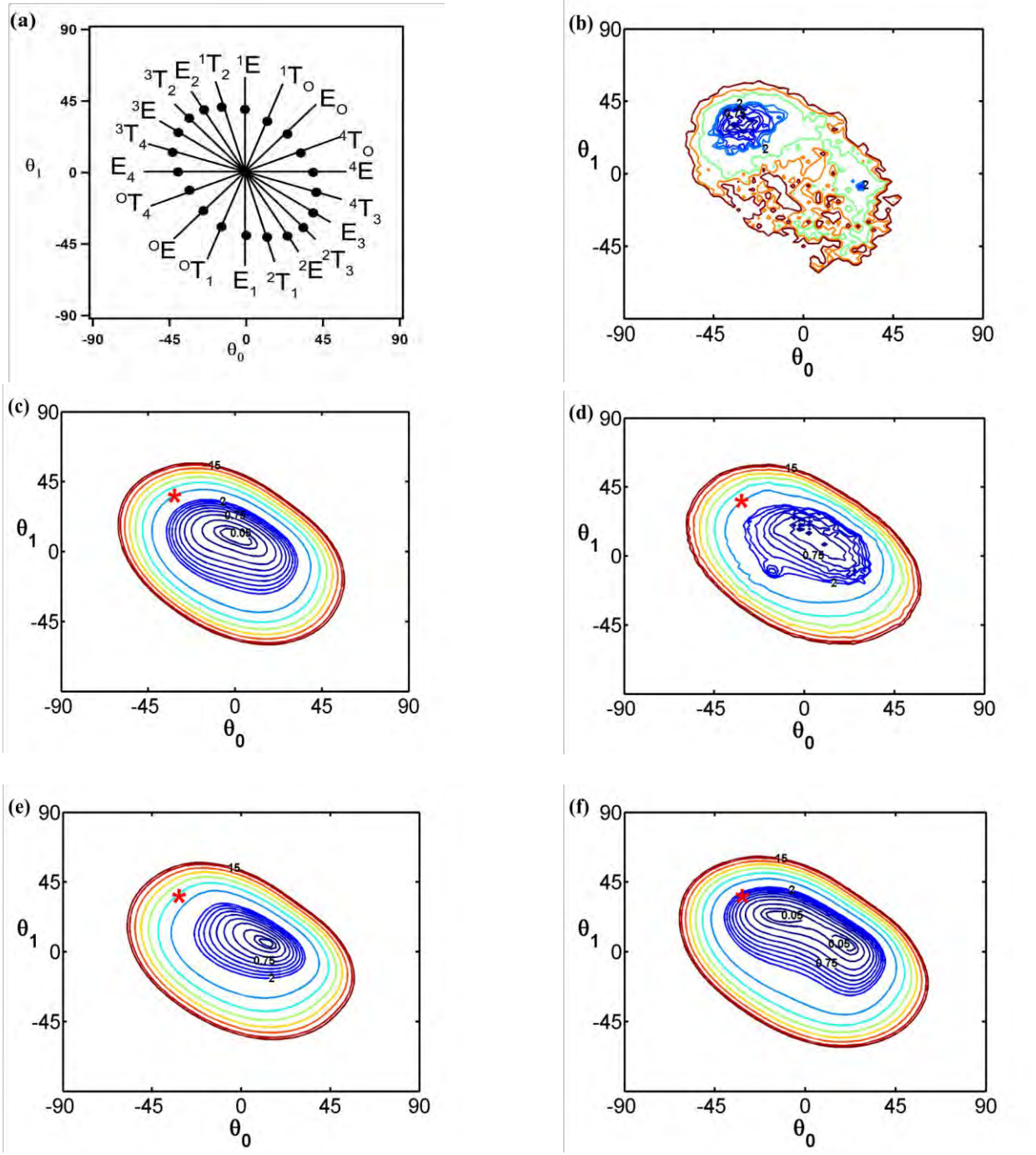


Figure 6.2: (a) Triangular tessellation pucker space for five-membered rings with canonical conformer coordinates shown as nodes. Ribofuranose free energy of puckering shown as two-dimensional contour plots for (b) HF/6-31G, (c) PM3CARB-1, (d) PM3^{MS} (e) AM1/d-PhoT and (f) AM1/d-CB1. Energy is mapped to color from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.05 kcal/mol to 0.1 kcal/mol, then from 0.1 kcal/mol to 2 kcal/mol in steps of 0.25 kcal/mol and every 2 kcal/mol thereafter. The HF global energy minimum (shown as red stars) is marked on each SE FES.

Here the global minimum (15°, 5°) is closest to a $^{4}T_{o}$ conformer with ^{1}E, $^{1}T_{o}$, E_{o}, $^{4}T_{o}$ and ^{4}E being 0.5 kcal/mol away at rim of the FES. NDDO methods therefore lead to models incapable of accurately describing distinct transitions states for furanose rings.

The AM1/d-CB1 (Figure 6.2f) global minimum (-12.5°, 22.5°) represents $^{1}T_{2}/E_{2}$ conformers that are in close proximity to the reference HF $^{3}T_{2}/^{3}E$ global minima conformers. A second minimum (20°, 5°) corresponding to a $^{4}T_{o}/^{4}E$ conformation with an energy of 0.01 kcal/mol can be accessed via the $^{1}T_{o}/E_{o}$ conformer (7.5°, 17.5°) of energy 0.09 kcal/mol of its global minimum. Approximately 0.8 kcal/mol is required to reach the planar conformation on the AM1/d-CB1 FES. The AM1/d-CB1 furanose 5-membered ring is more discriminate of canonical conformers and improves on NDDO methods commonly used to model carbohydrates.

6.1.2.2 Glucopyranose

For glucopyranose the free energy volumes are a function of (θ_0, θ_1, θ_2) with reference and rotatable planes chosen using a method of triangular tessellation as described previously.[12,14,17] The 38 canonical conformers for six-membered rings comprising chairs (C), boats (B), twists/skews (S), half-chairs (H), and envelopes (E) are shown in Figure 6.3(a). The free energy of glucopyranose as a function of reaction coordinates $W(\theta_0, \theta_1, \theta_2)$ is visualized in the three dimensions of the reaction coordinates representing the free energy in colour where low free energies (0 kcal/mol) are blue and very high free energies (8 kcal/mol) are red (Figure 6.3b-k). In each of the free energy landscapes there are two isovolumes. The first is an inner (turquois) surface at 3 kcal/mol (*~5kT*) representing pucker conformers observed at equilibrium dynamics. The second (outer) surface encases minimum energy pathways from the $^{4}C_{1}$ conformation (south pole), where all hydroxyl groups are equatorial, to the $^{1}C_{4}$ conformation (north pole), where all hydroxyls are axial. The lowest energy conformer predicted for all methods, other than AM1/d-PhoT, is $^{4}C_{1}$. At 3 kcal/mol AM1/d-PhoT shows the existence of skew boats and boats ($^{O,3}B$, $^{O}S_{2}$, $B_{2,5}$, $^{1}S_{5}$, $^{1,4}B$, $^{1}S_{3}$, $^{2,5}B$, and $^{5}S_{1}$) as well as the $^{3}H_{2}$ and ^{3}E conformers (Figure 6.3b).

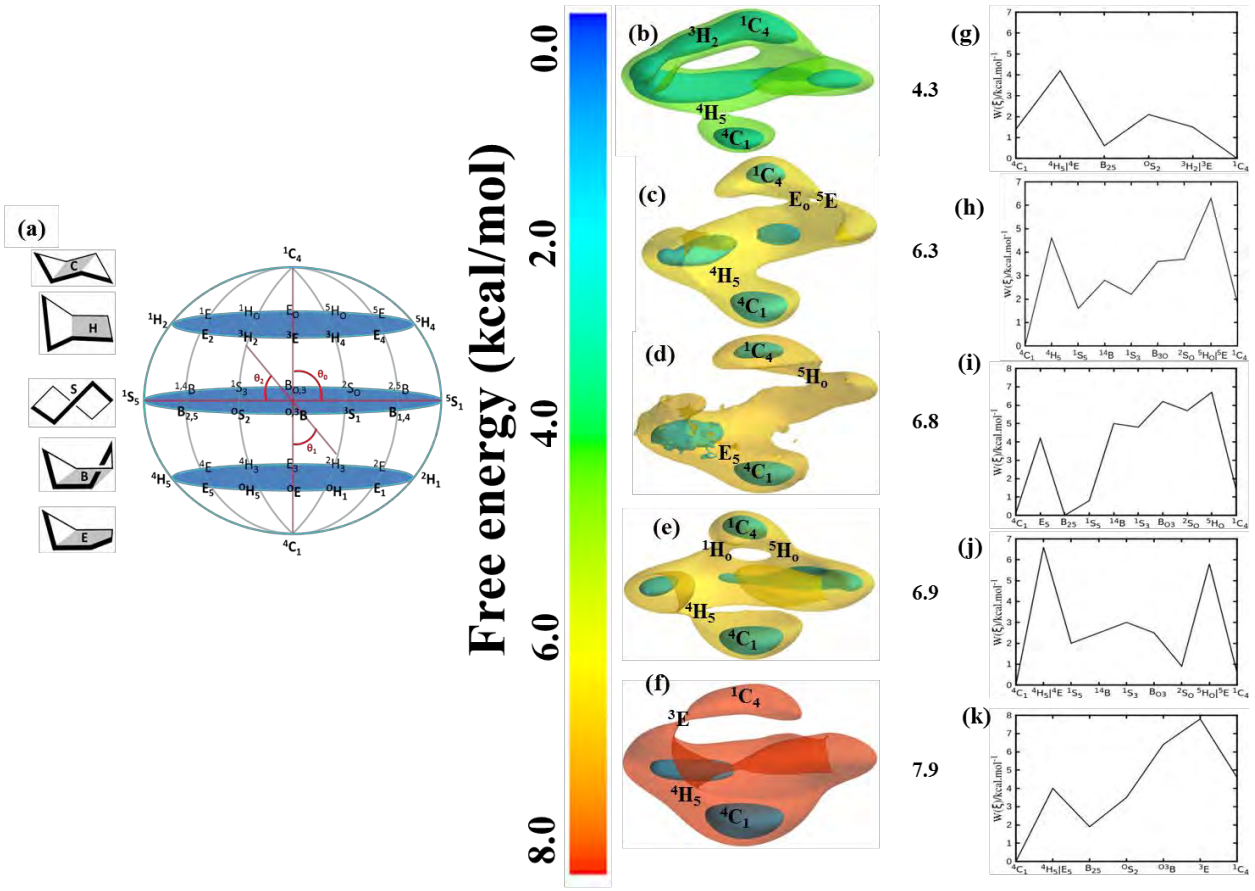


Figure 6.3: (a) Canonical conformers projected onto the triangular tessellated pucker coordinates (θ_0, θ_1, θ_2) for six-membered rings. The free energy $W(\theta_0, \theta_1, \theta_2)$ volumes for (b) AM1/d-PhoT, (c) AM1/d-CB1, (d) PM3^{MS}, (e) SCC-DFTB and (f) PM3CARB-1 are shown on color. The free energy values are mapped in color from 0 kcal/mol (blue) to 8 kcal/mol (red). The inner isosurface is at 3 kcal/mol and the outer isosurface indicates the minimum free energy to connect the $^1C_4$ and $^4C_1$ conformers which occurs at 4.3, 6.3, 6.8, 6.9 and 7.9 kcal/mol, respectively. The one-dimensional minimum free energy paths have been extracted from the free energy volumes and are shown for (g) AM1/d-PhoT, (h) AM1/d-CB1, (i) PM3^{MS}, (j) SCC-DFTB and (k) PM3CARB-1.

AM1/d-CB1, PM3^{MS}, PM3CARB-1 and SCC-DFTB provide more restricted minimum energy paths (Figure 6.3c-f) compared to AM1/d-PhoT. The barrier heights separating $^4C_1$ from $^1C_4$ for AM1/d-PhoT are the lowest (4.3 kcal/mol) while AM1/d-CB1, PM3^{MS}, SCC-DFTB and PM3CARB-1 have barrier heights at least 1kcal/mol are higher (6.3, 6.8, 6.9 and 7.9 kcal/mol, respectively).

Minimum free energy paths have been extracted and plotted as line diagrams (Figure 6.3g-l). In the absence of an *ab initio* computed free energy pucker volume the SCC-DFTB volume and derived minimum path ${}^{4}C_{1}\rightarrow{}^{4}H_{5}/{}^{4}E\rightarrow{}^{1}S_{5}/{}^{1,4}B\rightarrow{}^{1}S_{3}\rightarrow B_{O,3}\rightarrow{}^{2}S_{O}\rightarrow({}^{5}H_{O}/{}^{5}E$ or ${}^{5}H_{O})\rightarrow{}^{1}C_{4}$ (Figure 6.3j) is used as a reference conformational mechanism for the ${}^{4}C_{1}$ to ${}^{1}C_{4}$ ring pucker transition. We do this as we had previously established that SCC-DFTB, of any SE method, best models carbohydrate ring pucker.[14] AM1/d-CB1 directly matches the SCC-DFTB minimum free energy pathway (Figure 6.3h) from the ${}^{4}C_{1}$ to ${}^{1}C_{4}$ conformer although the energy profile differs between the two methods. In the AM1/d-CB1 case the *H* and *E* conformers populating the southern "tropic" commonly associated with TS structures in glycosyltransferase (GT) catalyzed reactions have a barrier height of 4 kcal/mol. Whereas the same SCC-DFTB computed ${}^{4}H_{5}/{}^{4}E$ conformers are more than 6 kcal/mol higher on the free energy pucker volume than AM1/d-CB1. Along the northern "tropic" the *H* and *E* conformers have a barrier height of more than 6 kcal/mol that is close to the 5.8 kcal/mol barrier predicted by SCC-DFTB.

The AM1/d-PhoT minimum path ${}^{4}C_{1}\rightarrow{}^{4}H_{5}/{}^{4}E\rightarrow B_{2,5}\rightarrow{}^{O}S_{2}\rightarrow{}^{3}H_{2}/{}^{3}E\rightarrow{}^{1}C_{4}$ (Figure 6.3g) initially puckers in the same way as SCC-DFTB and AM1/d-CB1 with a ${}^{4}H_{5}/{}^{4}E$ conformation barrier of 4 kcal/mol but the glucopyranose ring crosses the *B* and *S* conformers at the "equator" and transitions to the ${}^{1}C_{4}$ conformer very differently from the former two methods.

The $PM3^{MS}$ minimum path (Figure 6.3i) deviates perhaps the furthest from the SCC-DFTB both in its puckering as well as the very high barriers (6-7 kcal/mol) it passes through toward the ${}^{1}C_{4}$ conformer. Moreover, the global minima on the pucker free energy surface is the $B_{2,5}$ conformer. This is inconsistent with accepted stereo electronic understanding of glucopyranose 6-membered ring pucker.[20]

The PM3CARB-1 minimum path (Figure 6.3k) passes through a ${}^{4}H_{5}/E_{5}$ 4 kcal/mol barrier to get to a $B_{2,5}$ minimum. From $B_{2,5}$ the ring is transformed into ${}^{O}S_{2}$, ${}^{O,3}B$ and ${}^{3}E$ conformers with increasingly large pucker energy. The ${}^{1}C_{4}$ conformer (4.6 kcal/mol) is higher in energy than the boats and skew-boats in the PM3CARB-1 free energy volume.

Three-dimensional free energy volumes as well as extracted one-dimensional free energy paths for AM1, PM3 and RM1 are provided in Appendix B (Figure B3).

6.1.3 Proton transfer

A GT reaction (Scheme 6.1) proceeds with the aid of a catalytic acid/base. Amino acid residues within the GT catalytic domain (label C) undergo proton transfers with the saccharide. Given its importance in the catalytic process, the proton affinity of amino acids commonly found in the GT catalytic domain as well as in the binding sites of lectins were computed using AM1/d-CB1 as well as methods commonly used to model carbohydrates. In addition we also consider the interaction energies of a number of base pairs that are applicable to DNA, RNA and numerous enzymes. The proton affinities of 18 amino acids (illustrated in Figure B4 of Appendix B) were computed in two ways. The first involves the protonation of the nitrogen on the amino group of a neutral amino acid leading to a positively charged species. Here reference values were obtained from Gronert et al.,[21] as well as experimentally available data posted in the NIST database.[22,23] The second involves protonating the oxygen of the negatively charged carboxylate leading to a neutral amino acid. Here we used CBS-QB3 calculations computed using Gaussian 09,[24] that had been shown to be accurate,[8,25,26] to generate reference proton affinities values. In the case of nitrogen protonation geometry optimization calculations on G3MP2 geometries were computed while in the case of oxygen protonation geometry optimization calculations on the CBS-QB3 geometries were computed. All calculations were done using a modified version of the MOPAC7.0 software package. The PM3CARB-1, PM3^{MS}, AM1/d-PhoT and AM1/d-CB1 MUEs of both the N- and O-protonated amino acids are shown in Figure 6.4. AM1/d-PhoT has the smallest error for the protonation of nitrogen with a value of 3.5 kcal/mol (Table B4, Appendix B), with AM1/d-CB1 having a slightly larger error (MUE of 5.1 kcal/mol). PM3CARB-1 and PM3^{MS} are considerably inaccurate with errors of 74.0 and 27.5 kcal/mol, respectively.

AM1/d-CB1 and PM3^{MS} have the lowest errors 5.6 kcal/mol and 3.6 kcal/mol, respectively, for the anionic carboxylate protonation (Table B4, Appendix B) yielding values closest to the CBS-QB3 reference set. AM1/d-PhoT's performance is reasonable good with MUE of 6.3 kcal/mol. PM3CARB-1 however, has the extremely large MUE of 71.2. It is interesting to note that all SE methods that have too low average errors for the protonation of the amino acid nitrogen have poor performance for the protonation of oxygen. In contrast AM1/d-CB1 produces similar results for both species. This implies that the method is likely to model the amino acids within the GT catalytic domain well.

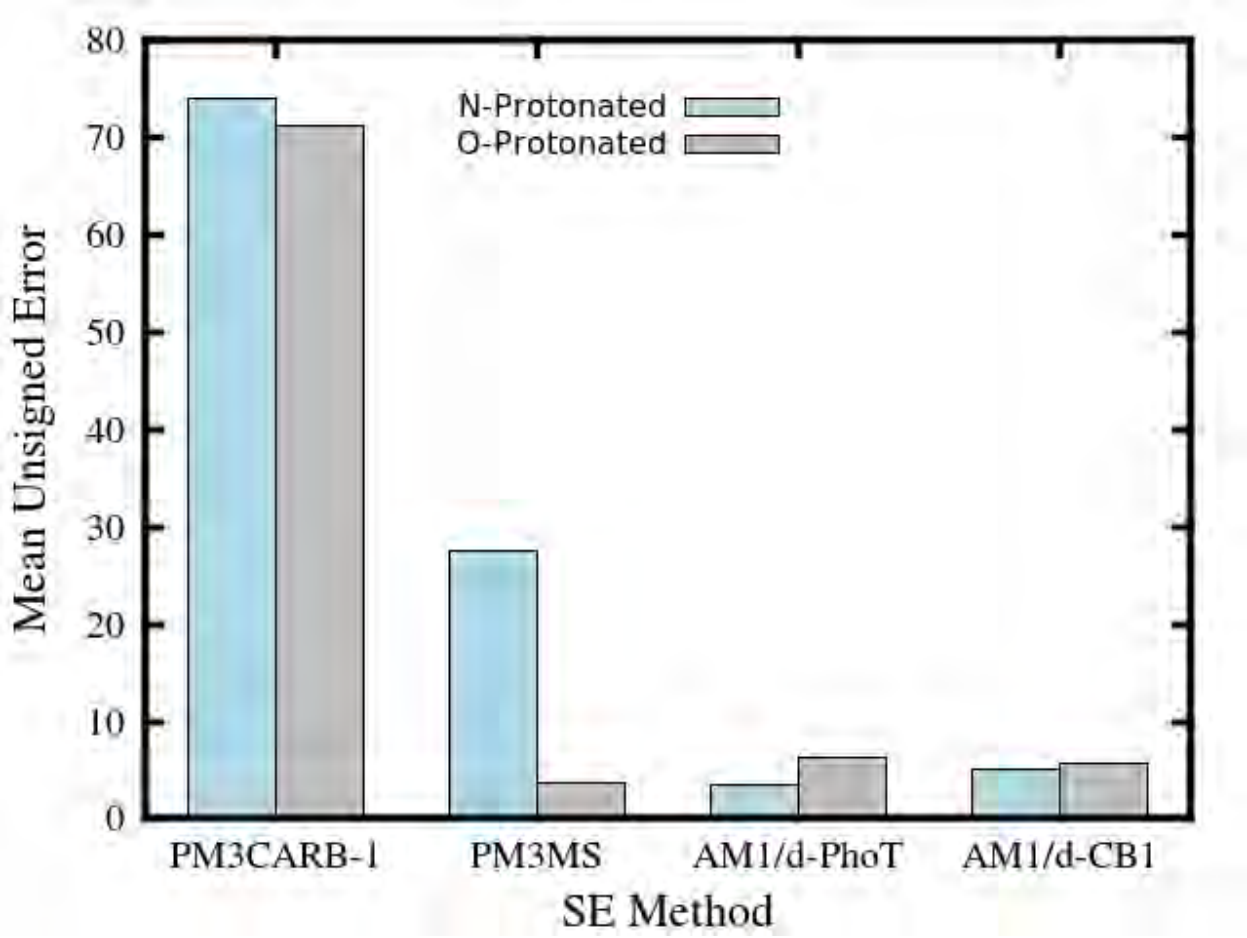


Figure 6.4: Geometry optimized mean unsigned errors for gas phase proton affinities of N- and O-protonated amino acids.

6.1.4 Nucleic acid base stacking and hydrogen bonding

A number of hydrogen bonded and stacked nitrogeneous base pairs were obtained from a benchmark database.[27] Both single point and geometry optimization calculations on the CCSD(T) optimized geometries using PM3CARB-1, $PM3^{MS}$, AM1/d-CB1 and AM1/d-PhoT were computed. The resulting interaction energies were then compared with the CCSD(T) values.

The distribution of computed interaction energies for hydrogen bonded and stacked base pairs (structures shown in Appendix B, Figures B5-B6) are shown in Figure 6.5. A comparison is made with DFT M06-2X energies computed by Hohenstein et al.[28] In general AM1/d-CB1, compares well with the CCSD(T) computed HB base pair interactions (MUE 3.63 kcal/mol)

although it slightly under estimates some of the interactions while all other NDDO methods significantly overestimates the hydrogen bonding (Figure 6.5a). Of the NDDO methods AM1/d-CB1 compares best (MUE 3.87 kcal/mol) with the CCSD(T) stacked base pairs data (Figure 6.5b). In both the hydrogen bonded and stacked base pairs calculations AM1/d-CB1 performs optimally, comparing with the CCSD(T) reference by at least two orders of magnitude better than PM3CARB-1, $PM3^{MS}$ and AM1/d-PhoT (Tables B5-B6).

None of the SE methods produce results as accurate as those of M06-2X/aug-cc-pVDZ[28] with errors of 1.61 and 0.98 kcal/mol for the hydrogen bonded and stacked complexes, respectively (Table B5). The comparison with of M06-2X with AM1/d-CB1 and other NDDO methods is not quite fair as the former includes corrections for dispersion and HB. In the case of the SE methods these interactions need to be accounted for empirically. As a future study we would like to include both empirical based dispersion and HB corrections to the AM1/d-CB1 Hamiltonian as we believe that it may produce better results.

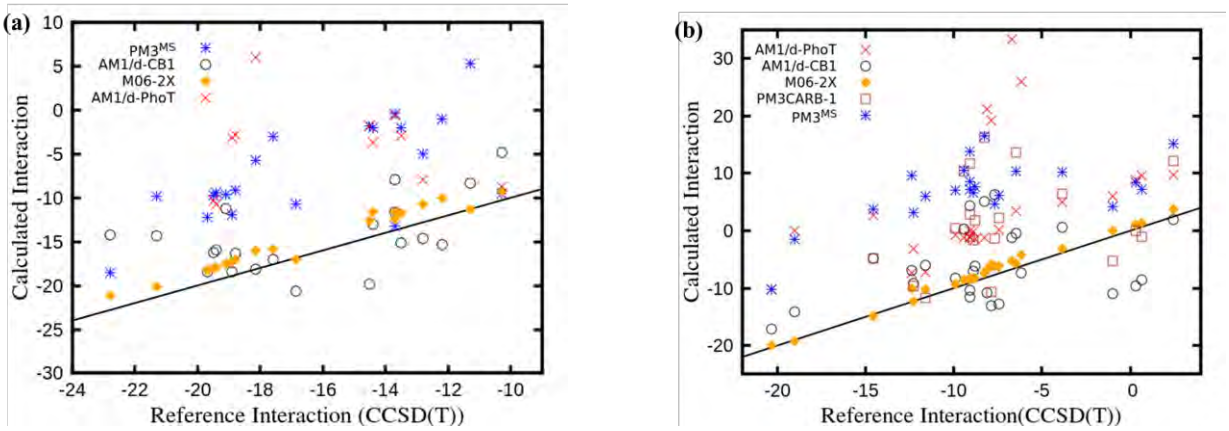


Figure 6.5: A comparison of NDDO and DFT/M06-2X gas phase interaction energies for (a) hydrogen bonded, and (b) stacked base pairs. The reference interaction energies are from CCSD(T) simulations. Interaction energies (kcal/mol) computed via geometry optimization.

Numerous single point calculations have also been performed, but not shown here (Appendix B, Figure B7 and Tables B7-B8).

6.1.5 Carbohydrate-aromatic π interactions

Amino acids with aromatic side chains, such as tryptophan, tyrosine, and phenylalanine form the basis of carbohydrate protein binding interactions and so are frequently found in protein active sites that recognize carbohydrates.[29] That adds a further dimension to the proton acceptor/donor properties to be accounted for in the glycoenzyme catalytic domains. Raju et al conducted a number of calculations[29,30] where they modelled the carbohydrate-aromatic interactions with high level *ab initio* computations. They used model complexes comprising carbohydrates interacting with toluene, p-hydroxytoluene, and 3-methylindole (analogues of phenylalanine, tyrosine, and tryptophan, respectively). Here we use the QM optimized carbohydrate-aromatic interacting complexes generated by Raju et al.[29,30] and conduct SE calculations on these complexes (Figure 6.6).

AM1/d-PhoT gives the smallest errors for both single point (2.89 kcal/mol) and geometry optimized (4.76 kcal/mol) structures (Table B9-B10, Appendix B). AM1/d-CB1 performs poorly with a MUE of 6.44 kcal/mol for both single point and geometry optimized interaction energies. For the complexes considered here the predominant interaction is that of dispersion and we expect a significant improvement, specifically for AM1/d-CB1, once an empirical based dispersion correction has been added and re-parameterized for the Hamiltonian. However, as mentioned above this is the subject of work that will be presented in a later publication.

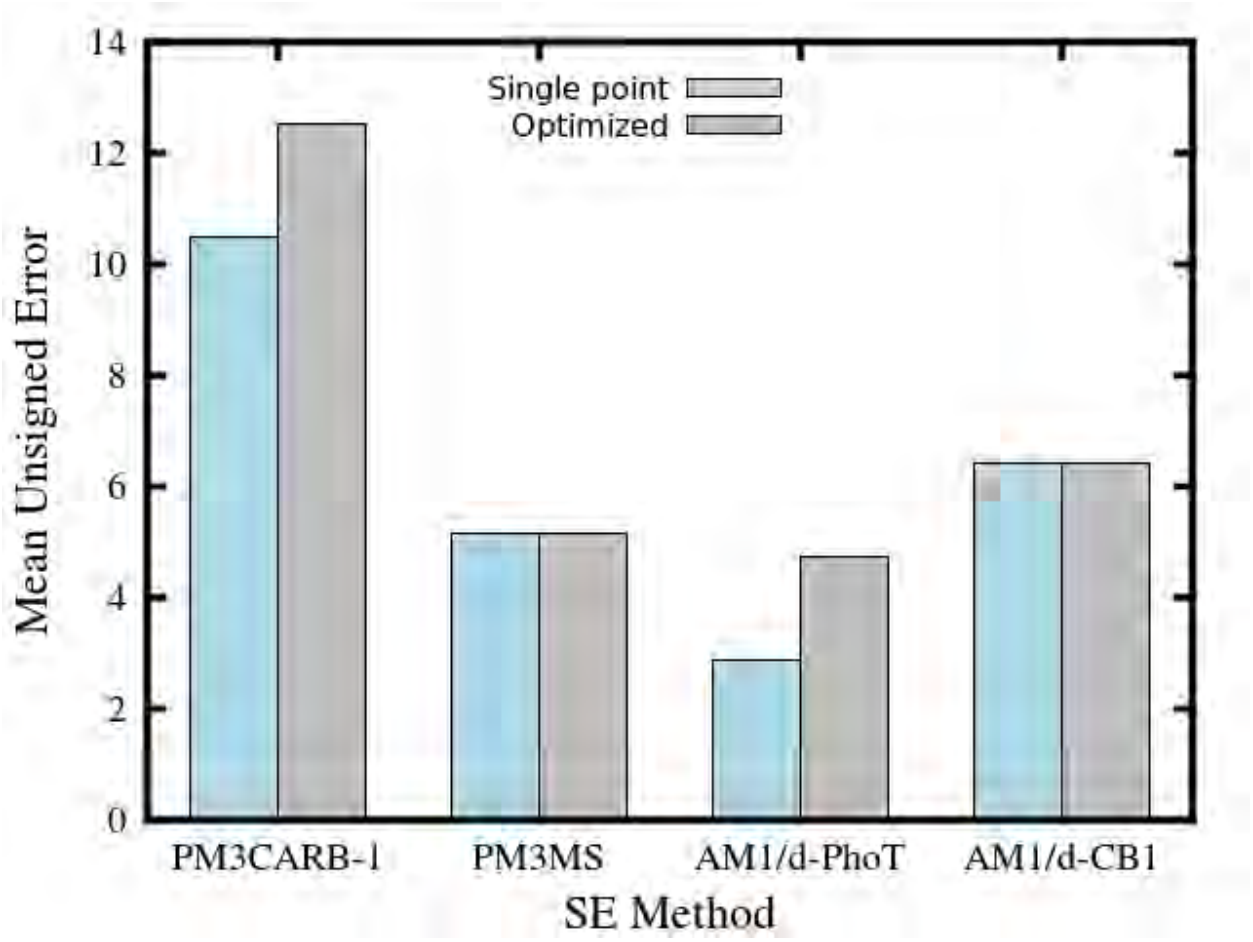


Figure 6.6: PM3CARB-1, PM3^{MS}, AM1/d-PhoT and AM1/d-CB1 MUEs computed for *ab initio* generated structures[29,30] of carbohydrate−aromatic π interactions (kcal/mol).

6.1.6 Glycosyltransferase reaction

Up until now we have evaluated AM1/d-CB1's performance for separate components that contribute to the accurate modeling of chemical glycobiological events particularly those central to a glycoenzymatic reactions. A typical glycosyltransferase reaction featuring an inversion or retention mechanism at the anomeric position requires a nucleoside phosphate, nucleoside diphosphate or lipid phosphate leaving group present as illustrated in Scheme 6.1, label F.

Recently a QM(DFT)/MM investigation into the substrate-assisted catalytic mechanism of O-linked N-acetylglucosamine (O-GlcNAc) transferase was reported by Tvaroška et al.[31] A schematic for the reaction that includes surrounding residues (UDP-GlcNAc, Val20, Ser21,

Ser22, His498, His558, Gln839, Lys842, Lys898, His901, His920, and three water molecules in the vicinity of UDP-GlcNAc) is given in Figure 6.7a.

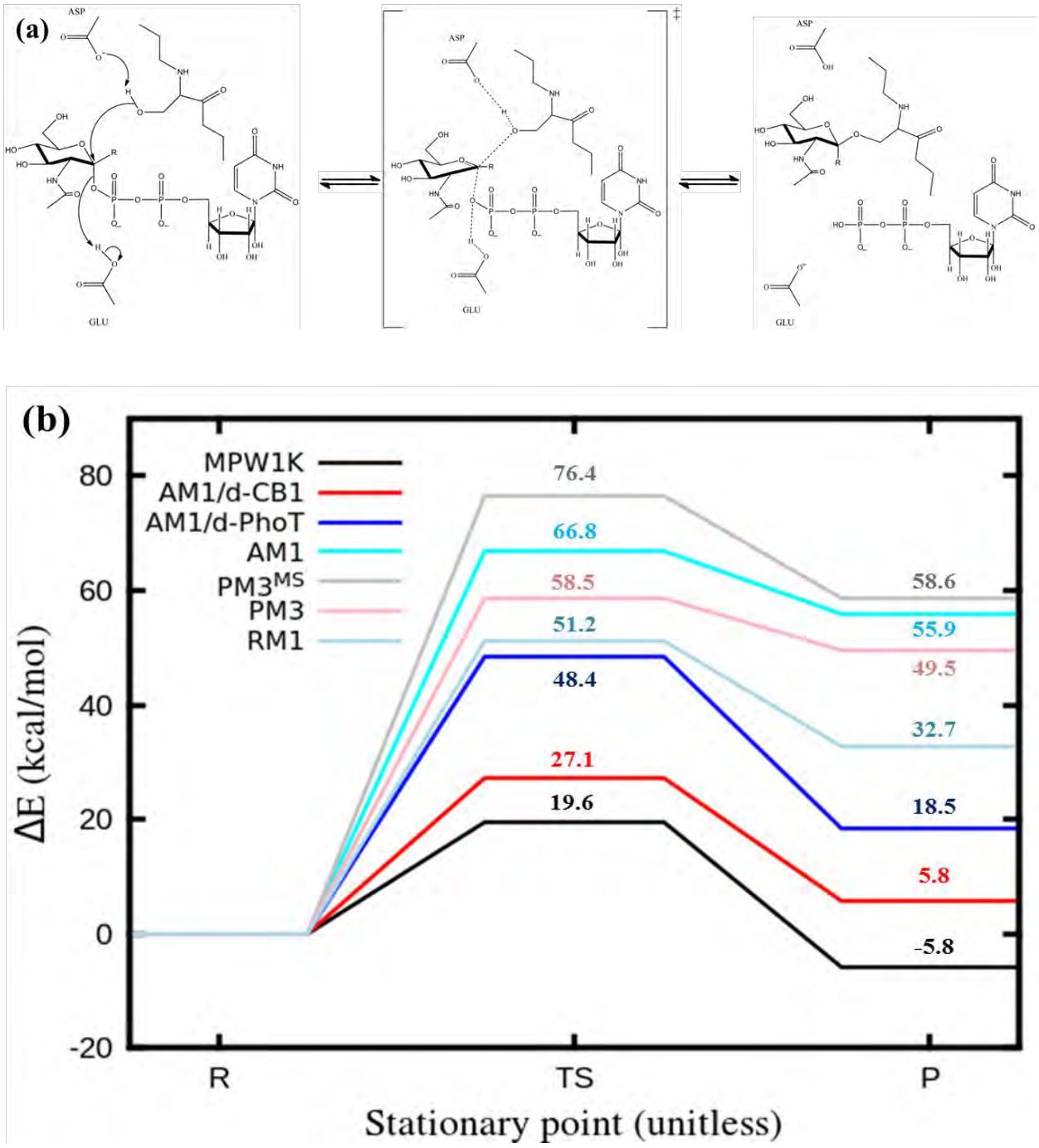


Figure 6.7: (a) Reaction scheme for enzymatic reaction catalyzed by uridine diphospho-N-acetylglucosamine polypeptide *β*-N-acetylaminyltransferase and (b) geometry optimized QM/MM 1D reaction profile energy traces. MPW1K profile was obtained from work by Tvaroška et al.[31]

We obtained the QM/MM geometries for the reactant (R), transition state (TS) and product (P) that were computed at the MPW1K level of theory using a combination of the 6-31G** and 6-31+G* basis sets,[31] as well as the OPLS force field for the classical region.[32] Single point QM/MM calculations of these structures were then performed using various SE methods (Table B11) contained within the CHARMM/MNDO97[16] interface. The MM was treated with the OPLS force field[32] in order to ensure that the calculations are in line with those performed by Tvaroška et al.[31] The energy barriers for the reaction are provided in Figure 6.7b. The energies obtained for the MPW1K level are 19.6 kcal/mol (TS) and -5.8 kcal/mol (P) relative to R. It is important to note that the QM region defined by Tvaroška et al.[31] consisted of a total of 198 atoms, which was too large for the MNDO97 software package. As such the SE QM/MM calculations needed to have a smaller QM region consisting of only 81 atoms (UDG-GlcNAc, Ser21 and His498). The barriers obtained from the optimized QM/MM calculations are provided in Figure 6.7b as well as Tables B11-B13 of Appendix B. AM1/d-CB1 surpasses all of the other NDDO methods in predicting both the TS and P barriers with energies of 27.1 kcal/mol and 5.8 kcal/mol, respectively. In addition the optimized structures do not differ much from those obtained by the MPW1K method, which is apparent from the RMSD provided in Table B12 (Appendix B). Here we find that for the 6-membered carbohydrate ring, which participates in the reaction, it is AM1 and AM1/d-CB1 that have the smallest RMSD's for reactants (0.012345 and 0.013110, respectively) and products (0.011491 and 0.012373, respectively). For the transition state AM1/d-CB1 surpassed all other methods resulting in an RMSD of 0.018694. None of the NDDO methods produce a change in the transition state pucker ($^4H_3$) even after a geometry optimization.

Comparing the bond lengths of the reaction coordinates we find that AM1, AM1/d-PhoT and AM1/d-CB1 produce the best bond lengths when compared to the MPW1K results (Table B13). Single point QM/MM calculations have also been performed but are not presented here (Table B11).

6.2 Conclusion

The applicability of the newly parameterized AM1/d-CB1 SE method to chemical glycobiology has been evaluated. The method gives accurate results for molecular structures of monosaccharides that are important in mammalian biology. AM1/d-CB1 also shows

considerably different behavior for both 5- and 6-membered ring puckering when compared to other NDDO type methods (AM1, PM3, PM3CARB-1, $PM3^{MS}$, RM1, and AM1/d-PhoT) producing more than just one minimum energy conformer for the ribofuranose (5-membered) ring. This addresses a common weakness exhibited by NDDO methods where the pucker free energy surface of five membered sugar rings display no stationary points that correlate with canonical conformers. Further NDDO methods favour flattened five membered sugar rings showing no canonical ring pucker preferences. This is unlike AM1/d-CB1 that discriminates between canonical pucker conformers and presents a global minimum distinctly shifted away from the flattened ring. In the case of glucopyranose the ring is no longer unrealistically flexible as is the case with AM1, PM3, and $PM3^{MS}$; instead AM1/d-CB1 yields a $^4C_1$ to $^1C_4$ ring pucker pathway that directly matches the SCC-DFTB $^4C_1$ to $^1C_4$ pathway. Comparing with SCC-DFTB is best since no *ab initio* free energy pucker volume benchmarks exist.

While the accurate modelling of carbohydrates by AM1/d-CB1 is an important measure of its performance it is critical to examine its performance when computing key properties of molecules often present in the glycan environment. An example of this is the amino acid proton affinities that are central to the acid/base catalytic mechanism commandeered by GTs and other glycoenzymes. Here AM1/d-CB1 (as well as AM1/d-PhoT) yields the most accurate proton affinities for the protonation of nitrogen amino group of a neutral amino acid. In the case of oxygen protonation AM1/d-CB1 (as well as $PM3^{MS}$) gives the most accurate results when compared to other NDDO type methods. Therefore overall we expect AM1/d-CB1 to reliably model the acid/base contributions to the glycosidase or glycosyltransferase catalytic mechanisms.

Another important feature of carbohydrate protein interactions is the role that aromatic groups play in the binding of glycans in glycoenzyme catalytic domains as well as in the recognition sites of proteins such as lectins. Here AM1/d-CB1 gives a poor performance, but it is believed that the method can be improved upon by including and re-parameterization an empirical dispersion correction.

The AM1/d-CB1 parameter set is by no means perfect although it achieves better performances for the range of chemical glycobiological important property calculations than other NDDO methods. It suffers from the same perennials shortcomings underlying the NDDO method regarding hydrogen bond and dispersion interaction modeling. Therefore we believe that

AM1/d-CB1 when coupled with specifically optimized post SCF semi empirical HB and dispersion corrections will be capable of delivering optimal semi-empirical performance for chemical glycobiological events. These corrections are currently under development.

Notwithstanding the current lack of hydrogen bond and dispersion corrections the overall performance of AM1/d-CB1 as a reliable SE method for chemical glycobiology applications is apparent when we used it to compute the reaction profile of a recently studied GT. Here it gave the lowest transition state and product barrier compared with DFT (MPW1K) computations.

6.3 References

(1) Govender, K. K.; Gao, J.; Naidoo, K. J. *J. Chem. Theory Comput.* **2014**, Submitted.
(2) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(3) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.
(4) Stewart, J. J. P. *J Comput Chem* **1989**, *10*, 221.
(5) McNamara, J. P.; Muslim, A.-M.; Abdel-Aal, H.; Wang, H.; Mohr, M.; Hillier, I. H.; Bryce, R. A. *Chem. Phys. Lett.* **2004**, *394*, 429.
(6) Mane, J. Y.; Klobukowski, M. *Chem. Phys. Lett.* **2010**, *500*, 140.
(7) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101.
(8) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(9) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. J. *J. Phys. Chem. B* **2001**, *105*, 569.
(10) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
(11) Strümpfer, J.; Naidoo, K. J. *J. Comp. Chem.* **2010**, *31*, 308.
(12) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243.
(13) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.
(14) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.
(15) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(16) Thiel, W. University of Zurich, Zurich, Switzerland, 1998.
(17) Hill, A. D.; Reilly, P. J. *J. Chem. Inf. Model.* **2007**, *47*, 1031.
(18) Khalili, P.; Barnett, C. B.; Naidoo, K. J. *J. Chem. Phys.* **2013**, *138*, 184110.
(19) Taniguchi, N.; Honke, K.; Fukuda, M. *Handbook of Glycosyltransferase and Related Genes.*; Springer: Tokyo, 2002.
(20) Stoddart, J. F. *Stereochemistry of Carbohydrates*; John Wiley & Sons, Inc.: New York, 1971.
(21) Gronert, S.; Simpson, D. C.; Conner, K. M. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2116.

(22) Hunter, E. P.; Lias, S. G. *Nist Standard Reference Database Number 69*; Mallard, W. G., Linstrom, P. J., Eds. National Institute of Standards and Technology (http://webbook.nist.gov): Gaithersburg MD, 2008.
(23) Linstrom, P.; Mallard, W. *NIST Chemistry WebBook, http://webbook.nist.gov/chemistry;* NIST Standard Reference Database Number 69: National Institute of Standards and Technology, Gaithersburg MD, 2003.
(24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, Inc., Wallingford CT: 2009.
(25) Moser, A.; Range, K.; York, D. M. *J. Phys. Chem. B* **2010**, *114*, 13911.
(26) Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.
(27) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
(28) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. *J. Chem. Theory Comput.* **2008**, *4*, 1996.
(29) Raju, R. K.; Ramraj, A.; Vincent, M. A.; Hillier, I. H.; Burton, N. A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6500.
(30) Raju, R. K.; Ramraj, A.; Hillier, I. H.; Vincent, M. A.; Burton, N. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3411.
(31) Tvaroška, I.; Kozmon, S.; Wimmerová, M.; Koča, J. *J. Am. Chem. Soc.* **2012**, *134*, 15563.
(32) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc* **1999**, *121*, 4827.

7. AM1*-CB1: A diatomic core based semi-empirical method for chemical glycobiology systems

The new phosphorus parameters of AM1-CB1 are given and their accuracy in predicting properties that were utilized for the training set are evaluated. A test is also performed on a phosphoric reaction that is prominent in glycosyltransferases as an additional evaluation of the methods accuracy when applied to the field of chemical glycobiology.*

7.1 Background

Before considering the results produced for AM1*-CB1, from both a development and application point of view, there are a few important aspects which need to stressed.

It has been well established (Chapters 2 and 3) that methods such as AM1/d-PhoT[1] and AM1*[2] utilize the standard AM1 Hamiltonian's theory and approximations when treating systems which possess only *s*- and *p*-orbitals (organic systems). For molecular systems that possess *d*-orbitals both methods make use of the standard MNDO/d[3,4] approximations (Chapter 3, section 3.2.4). However, the manner in which the repulsive interactions of *d*-orbital systems are treated gives rise to the very important difference between these two methods. In such a case AM1/d-PhoT makes use of a modified AM1 core-core repulsion (eq. 3.81), while AM1* has a core-core repulsion that is depended on a set of bond specific parameters (eq. 3.79). As we have stated previously (Chapter 4) AM1/d-PhoT was the building block used to obtain the parameters for AM1/d-CB1 (Chapters 5 and 6). Similarly AM1* was used as the starting point to obtain parameters for AM1*-CB1. Since organic systems are treated the same for both AM1/d-PhoT and AM1* it was decided that the two newly parameterized methods (AM1/d-CB1 and AM1*-CB1) would possess the same parameters for H, C, N and O. In this way we ensure that the accurately predicted properties of AM1/d-CB1 (carbohydrate ring puckering, amino acid proton affinities and base pair interactions) are transferred to AM1*-CB1. Therefore, the only parameters that required optimization during the parameterization of AM1*-CB1 were those of phosphorus. In the sections that follow we shall look at the results obtained for AM1*-CB1,

focusing specifically on systems in our training set that possess phosphorus. We shall also look at some applications related to phosphates, which are important to chemical glycobiology.

7.2 Results and Discussion

As stated above the AM1*-CB1 parameters for H, C, N, and O are identical to those of AM1/d-CB1 (provided in Chapter 5). Following the procedure outlined in Chapter 4 an optimum set of phosphorus parameters were obtained for AM1*-CB1 and these are listed in Table 7.1 along with the original AM1* parameters that were used as the starting point.

Table 7.1: Optimized AM1*-CB1 parameters for Phosphorus along with original AM1* parameters for comparison.

	Phosphorus	
Parameters	AM1*-CB1	AM1*
U_{ss}	-48.814007	-45.6707151
U_{pp}	-32.836209	-35.2098162
U_{dd}	-22.255710	-23.6885421
ζ_s	2.126708	2.0894704
ζ_p	2.113663	1.9476331
ζ_d	1.158573	1.2697580
β_s	-10.744109	-10.3868963
β_p	-9.864607	-10.7694019
β_d	-4.966701	-4.9129999
α	1.913469	1.8232300
G_{ss}	11.970620	10.9221093
G_{pp}	8.093306	8.5031975
G_{sp}	5.085241	5.6174929
G_{p2}	7.825399	7.8119356
H_{sp}	0.815045	0.7461127
ζ_{sn}	1.520183	1.6351391
ζ_{pn}	1.018001	0.9773978
ζ_{dn}	0.952572	0.8744020
ρ_{core}	1.181260	1.2437106
F^0_{sd}	12.639316	11.6055655
G^2_{sd}	11.851646	12.9748658
α_{ij}		
P−H	1.864491	1.7054944
P−C	0.995627	1.7662992
P−O	1.857234	2.1041690
δ_{ij}		
P−H	1.013624	1.0906700
P−C	2.225642	1.0734607
P−O	1.596329	1.6352693

From the original AM1* paper[2] it shall be seen that diatomic parameters are also reported for P-N, P-F, P-P, P-S, P-Cl, and P-Mo interactions. These were not included in this work as none of these interactions were present within the molecules used for the training set. Since various *sp* based SE methods have been discussed in Chapters 5 and 6, in the sections that follow we shall focus on SE methods which make use of an *spd* basis for the hypervalent phosphorus.

7.2.1 Gas phase proton affinities

Protonation reactions are among the most important in chemistry and biology. This is the first step in many fundamental chemical rearrangements and in most enzymatic reactions. The proton affinity (PA) is the negative of the enthalpy change at standard conditions. Various studies have shown that accurate prediction of proton affinities with high-level QM simulations is possible.[5-13] As with the previous chapter (Chapter 5) the phosphates in our training set were grouped into molecular subclasses. A summary of these subclasses is provided in Table C1 of Appendix C. Tables C2-C3 provides a comparison of calculated PAs with those obtained from high level (M06-2X/6-311++G(3df,2p)) simulations. We would have made use of experimentally determined data for the PA, but due to the lack thereof we decided to use data from high level quantum simulations.

Figure 7.1 provides a breakdown of the mean unsigned errors (MUEs) obtained for the various phosphates used during parameterization. For the individual phosphate molecules AM1/d-PhoT surpasses all other methods producing a MUE of 3.4 kcal/mol (Table C2). AM1*-CB1 gives a MUE of 6.8 kcal/mol, that is two orders in magnitude larger than AM1/d-PhoT.

For the proton affinity prediction of carbohydrate-phosphate species we find that AM1*-CB1 produces results that are very similar to AM1/d-CB1. For β-glucose-phosphate it is AM1/d-PhoT that produces the best result (smallest error) with a MUE of 0.54 kcal/mol. In the case of β-ribofuranose-phosphate AM1*-CB1 has a MUE of 0.60 kcal/mol, which is close to that of AM1/d-CB1 (0.69 kcal/mol). With α-glucopyranose-phosphate all SE methods produce an unsigned error that is very similar as seen in Figure 7.1 (grey bar). Finally for α-ribofuranose-phosphate it is AM1*-CB1 that produces the smallest MUE of 0.38 kcal/mol. This is more than two orders in magnitude smaller than errors produced by the other SE methods.

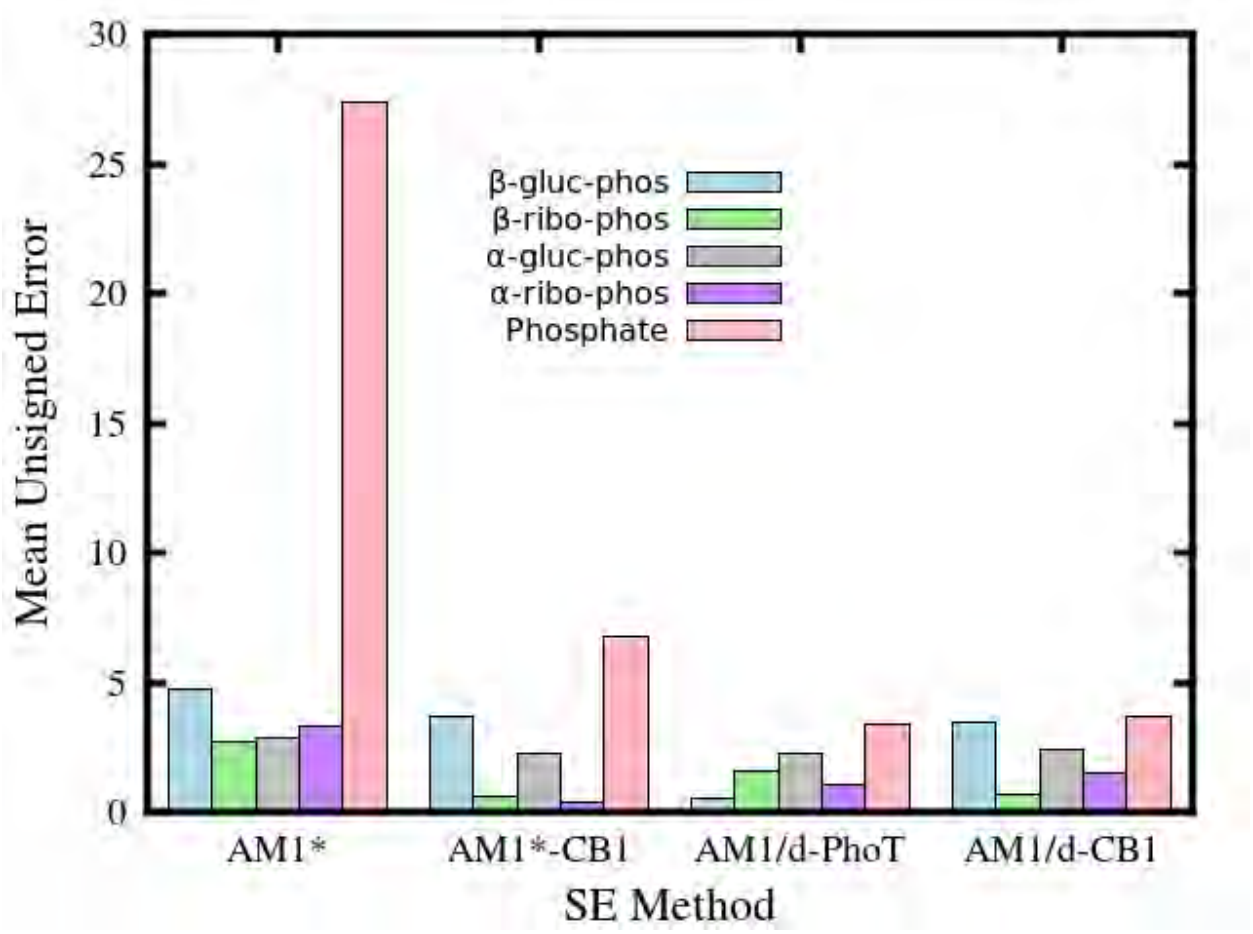


Figure 7.1: Mean unsigned errors for gas phase proton affinities of phosphates and carbohydrate chair phosphorylated conformers.

Despite the smaller errors obtained with AM1/d-PhoT for the carbohydrate-phosphate systems we know from previous chapters (Chapter 6) that the method is incapable of getting a fundamental aspect, present in glycobiological systems, modelled correctly (carbohydrate ring pucker). Therefore, the fact that the method produces smaller errors for the *β*-glucose-phosphate entails that the method is capable of getting proton acceptance/donation predicted correctly, but will have this occur at incorrect geometrical configurations. In the case of AM1*-CB1 the proton may not be accepted or donated as readily as AM1/d-PhoT (for *β*-glucose-phosphate), but since the organic parameters correspond to those of AM1/d-CB1, the method will predict the geometry of the glycobiological system correctly.

7.2.2 Dipole moments

A number of DFT (M06-2X) dipole moments were utilized during the course of parameterization, the results of which are provided in Table C4 (Appendix C). AM1* and AM1*-CB1 seems to outperform all other methods (Figure 7.2). In the case of the phosphates AM1* produces a MUE of 0.21 Debye, while AM1*-CB1 has a similar error of 0.24 Debye. AM1/d-PhoT has an error that is two orders in magnitude larger (0.47 Debye). Looking closely at the results produced in Table C4 we find that in the case of AM1*-CB1 the largest error comes from a species that is not important to chemical glycobiology ($PO(OCH_3)_3$). This species was included in the training set in order to try and make the new parameter set more robust and transferable to other chemical systems.

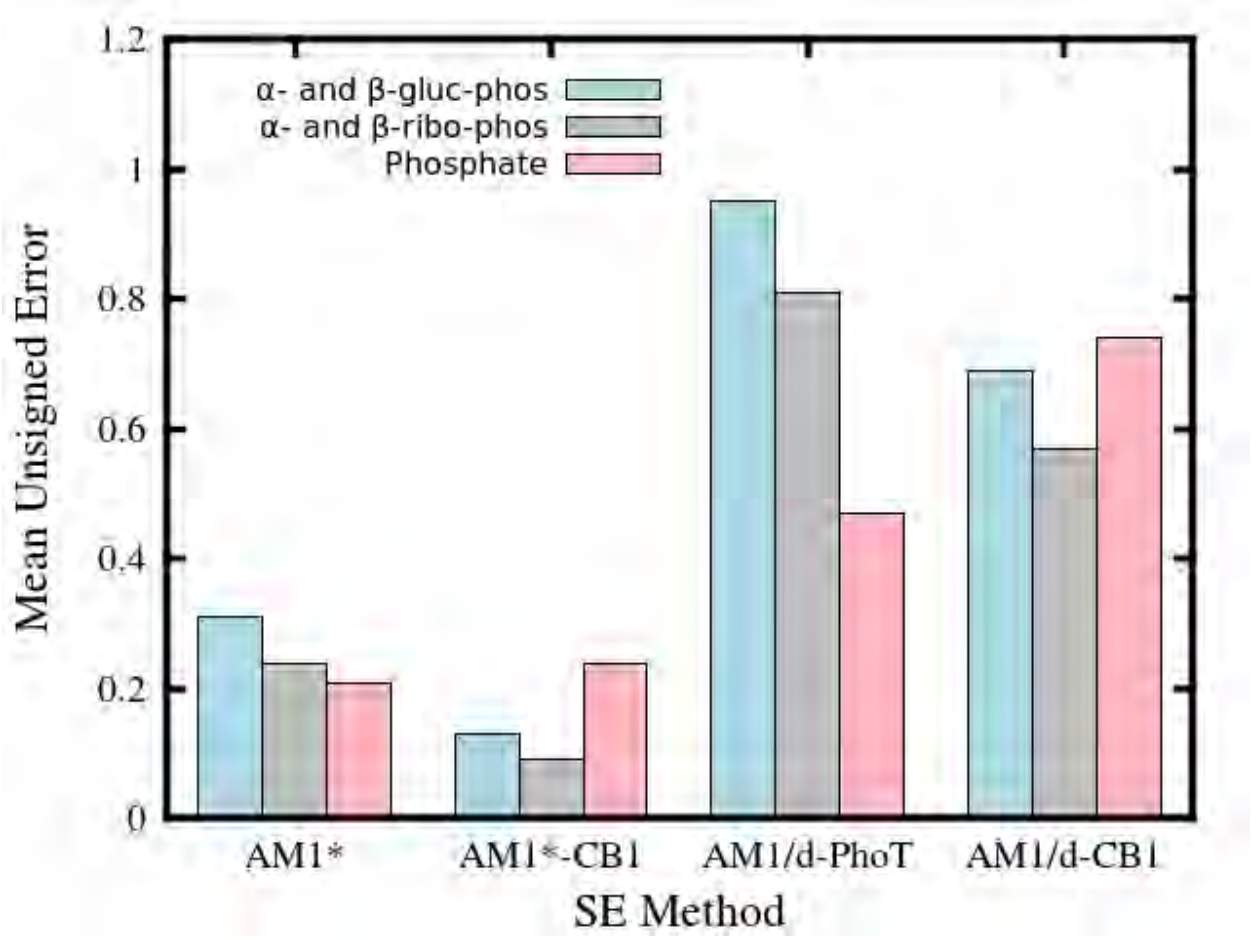


Figure 7.2: Mean unsigned errors for dipole moments of phosphates and carbohydrate chair phosphorylated conformers.

For the carbohydrate-phosphate systems AM1*-CB1 has MUEs of 0.13 and 0.09 Debye for the glucopyranose-phosphate and ribofuranose-phosphate, respectively, while AM1/d-CB1, AM1* and AM1/d-PhoT have errors that are substantially larger (Figure 7.2).

7.2.3 Ionization potential

A set of DFT based ionization potentials (IPs) were generated for various phosphoric species present in the training set and these are provided in Figure 7.3. The results (Table C5) show that all species considered are "stable" in the radical cationic forms with only positive ionization potentials being produced.

For the phosphoric systems AM1/d-PhoT produces the smallest MUE of 0.34 eV. AM1/d-CB1 and AM1*-CB1 give MUEs that are two to three orders in magnitude higher than AM1/d-PhoT (0.61 and 0.87 eV, respectively). The MUE for AM1* is the largest from all the methods with a value of 1.60 eV. It is clear that a significant improvement in IP has been achieved with AM1*-CB1 as the method has an error that is at least two orders in magnitude smaller than that of AM1*.

For the carbohydrate-phosphate systems AM1/d-PhoT once again produces the smallest MUE of 0.68 and 0.69 eV, for glucopyranose-phosphate and ribofuranose-phophate, respectively. AM1/d-CB1 and AM1*-CB1 have similar errors for glucopyranose-phosphate with MUEs of 1.18 and 1.04 eV, respectively. The corresponding errors for ribofuranose-phosphate are 1.02 and 0.91 eV, respectively. Extensive parameter optimization and refinement has been conducted on the phosphorus parameters of AM1*-CB1 and in some cases the method produced the smallest MUE for the carbohydrate-phosphate systems, while at the same time giving poor results (largest MUE) for all other properties considered in this work. Therefore, the IPs of the carbohydrate-phosphate systems is more dependent on the organic parameters of the carbohydrate and less dependent on the phosphorus parameters. However, a re-optimization of the organic parameters (H, C and O) would result in a poor pucker description (this has been tested), a property that is of significant importance in chemical glycobiology.

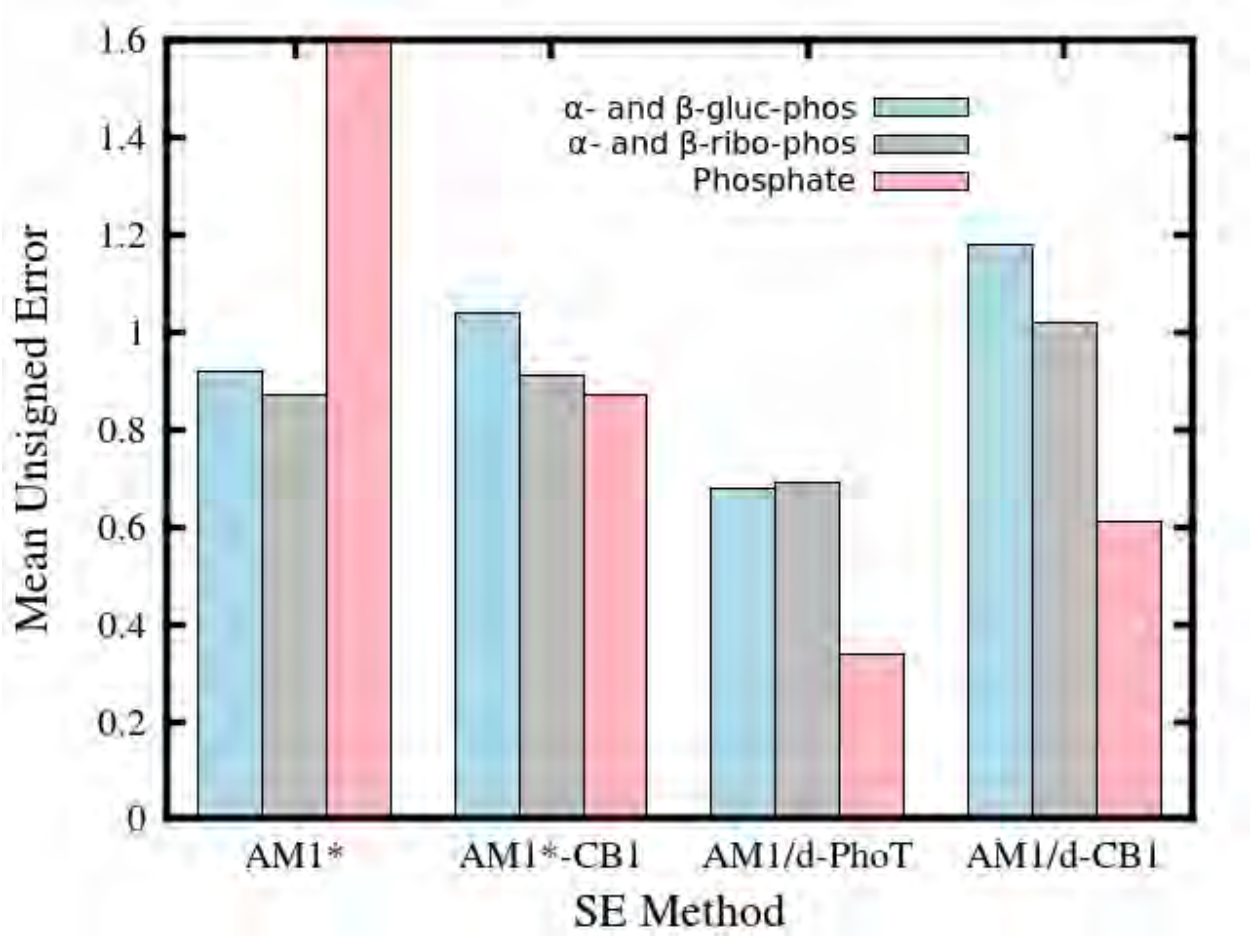


Figure 7.3: Mean unsigned errors for ionization potentials of phosphates and carbohydrate chair phosphorylated conformers.

7.2.4 Interaction energies

The interaction energies for various bimolecular complexes were computed using M06-2X/6-31+G(df) level of theory and basis set. A larger basis set (6-311++G(3df, 2p)) in this case produced substantially underestimated energies. Table 7.2 provides the interaction energies for various hydrogen bond dimers used in the training set. Results for *non*-phosphorus hydrogen bond dimers are emitted from here as their results are identical to those obtained for AM1/d-CB1 (Chapter 5, Table 5.3). The results clearly show that AM1/d-PhoT and AM1/d-CB1 produce results that are closest to those generated by DFT with MUEs of 1.40 and 1.60 kcal/mol, respectively. AM1*-CB1 has an error that is over 10 orders in magnitude higher than those of AM1/d-PhoT or AM1/d-CB1. A problem that was encountered during optimization of this property was that once interaction energies started to become more accurate (errors in line with

those produced by AM1/d-PhoT and AM1/d-CB1) there was in inversion experienced by the minimum energy conformers predicted for both the proton affinities and heats of formation of all carbohydrate phosphate systems, i.e. the minima predicted by AM1*-CB1 was different to those predicted by DFT ($^4C_1$ conformer instead of $^1C_4$ conformer). One will also note that the overall error of AM1*-CB1 is larger than that of AM1* and after extensive optimization this was a property that was sacrificed in order to obtain better results for other properties that were optimized during the parameterization.

Table 7.2: Theoretical interaction energies of selected molecules used in parameterization (kcal/mol)

Molecule	Reference	Error			
	DFT[a]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
$H_2O{:}PO_3^-$	-15.90	15.06	28.96	0.46	-0.77
$H_2O{:}H_2PO_4^-$	-18.20	17.59	29.47	-0.19	-2.41
$H_2O{:}HPO_4^{2-}$	-33.27	25.05	29.77	3.56	-1.62
MSE (vs DFT)		**19.23**	**29.40**	**1.28**	**-1.60**
MUE (vs DFT)		**19.23**	**29.40**	**1.40**	**1.60**

[a] DFT interaction energies where computed with M06-2X/6-31+G(df). All errors are computed as $\Delta H_{int}^{calc} - \Delta H_{int}^{DFT}$.

7.2.5 Heat of formation (ΔH_f)

For the heats of formation experimental values were used as target values for all phosphate systems (Table C6), while M06-2X calculations had to be used for all molecules that did not possess experimental data (carbohydrate-phosphate anomeric configurations). Figure 7.4 provides the MUEs obtained for the different phosphate based molecular subclasses. For the pure phosphate based molecules AM1*-CB1 produces a MUE (13.3 kcal/mol) that is very similar to AM1* (14.6 kcal/mol), while AM1/d-CB1 produce the smallest MUE of 9.6 kcal/mol.

For the carbohydrate-phosphate systems AM1*-CB1 has mixed success for the various anomers. With β-D-glucopyranose-phosphate the MUE for AM1*-CB1 is 2.76 kcal/mol, the smallest of all methods considered in this work. For the β- anomer of ribofuranose AM1*-CB1 and AM1/d-CB1 produce similar results with values of 1.87 and 1.44 kcal/mol, respectively. In the case of the α- anomers of both glucopyranose and ribofuranose AM1/d-PhoT yields the lowest errors of 1.48 and 1.07 kcal/mol, respectively. However for α-D-ribofurnaose-phosphate AM1*-CB1 and AM1/d-CB1 have errors that are very similar to that of AM1/d-PhoT with values of 1.54 and 1.18 kcal/mol, respectively.

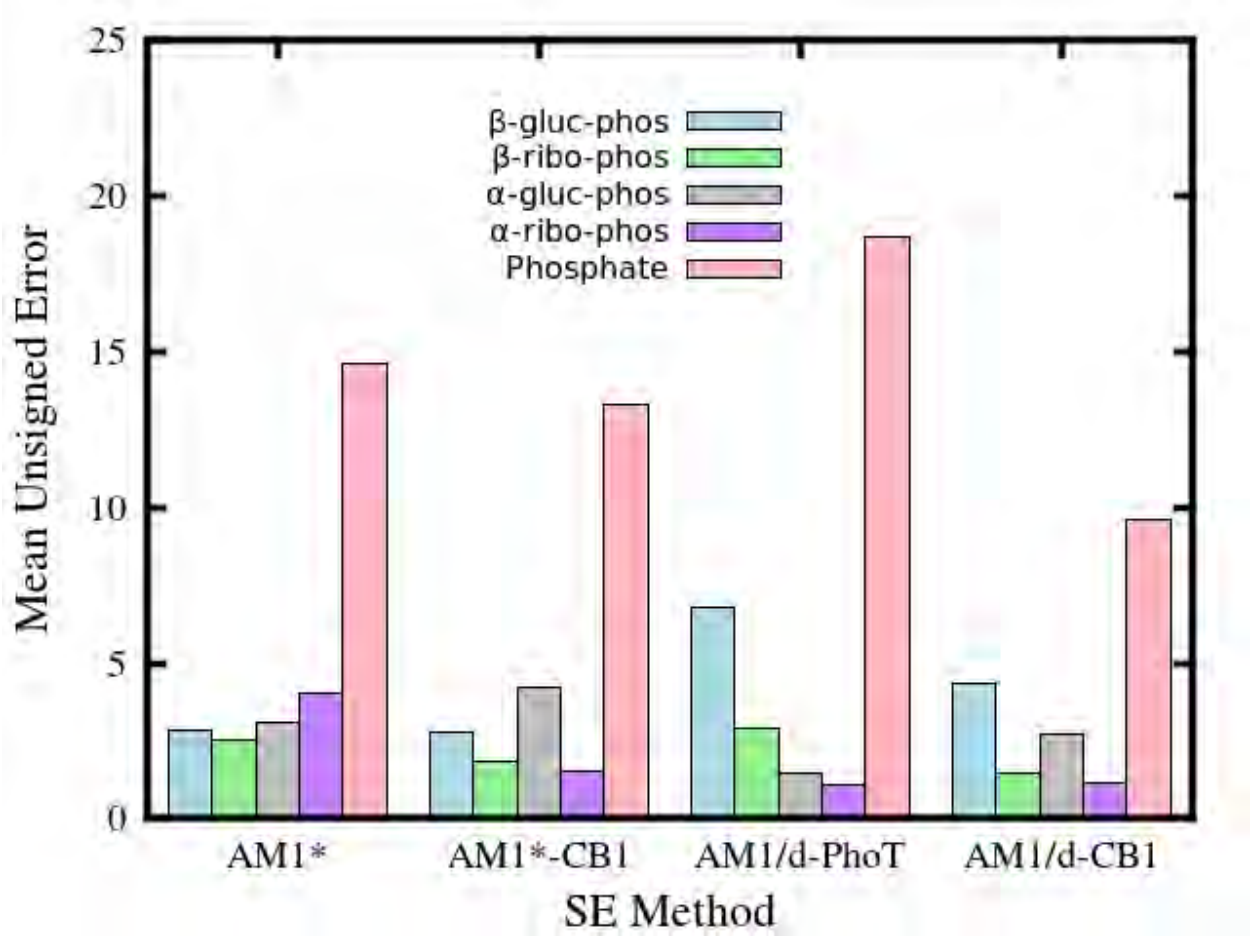


Figure 7.4: Mean unsigned errors for heats of formation of phosphates and carbohydrate chair phosphorylated conformers.

7.2.6 Phosphoric reactions

Having established the set of parameters for AM1*-CB1 the next step involves testing the new parameters on a test set of molecules which will closely mimic interactions which are experienced in glycosyltransferases. We have already illustrated the good performance of AM1/d-CB1 in Chapter 6 for various organic systems, such as carbohydrates (structure and ring pucker), amino acids (proton transfer), and amino acid base pairs (interaction energies). Since AM1*-CB1 possess the same H, C, N, and O parameters as AM1/d-CB1 the results for the two methods would be identical for the organic systems presented in Chapter 6. For phosphoric reactions, however, the two methods are expected to produce a different set of results since the phosphate follows a different SE framework and a different set of parameters when utilizing AM1/d-CB1 or AM1*-CB1.

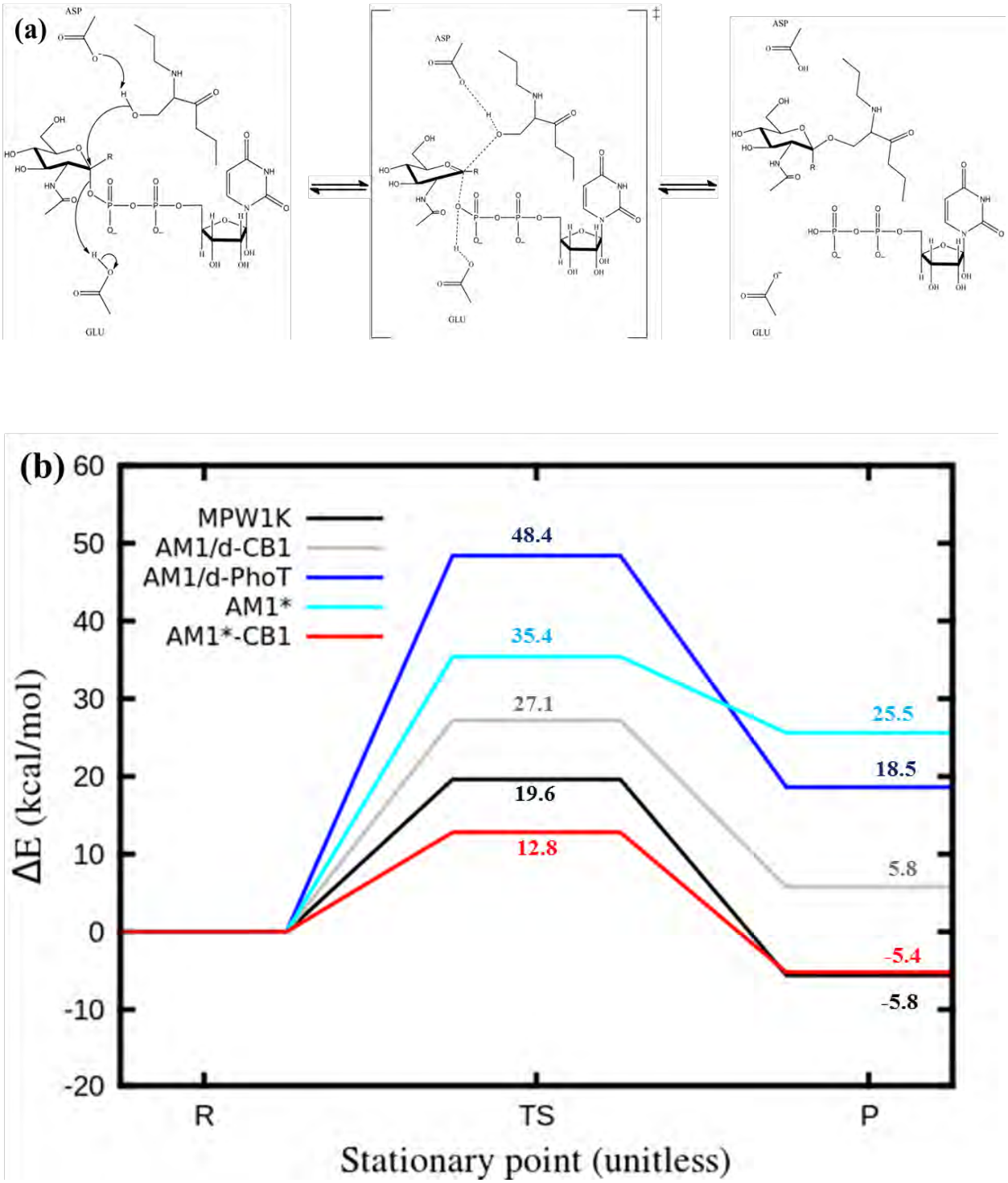


Figure 7.5: (a) Reaction scheme for enzymatic reaction catalyzed by uridine diphospho-N-acetylglucosamine polypeptide *β*-N-acetylaminyltransferase and (b) geometry optimized QM/MM 1D reaction profile energy traces. MPW1K profile was obtained from work by Tvaroška et al.[14]

The reaction which was utilized in Chapter 6 (section 6.1.5) is presented here once again. Table C8 (Appendix C) provides the energies obtained for the reaction that was tested in this work. A possible means to determine the accuracy of SE methods in modeling such a reaction would be to make use of QM/MM studies. Tvaroška et al.[15] has successfully modelled the substrate-assisted catalytic mechanism of O-linked N-acetylglucosamine (O-GlcNAc) transferase with the aid of QM(DFT)/MM studies. A schematic for the reaction that includes surrounding residues (UDP-GlcNAc, Val20, Ser21, Ser22, His498, His558, Gln839, Lys842, Lys898, His901, His920, and three water molecules in the vicinity of UDP-GlcNAc) is given in Figure 7.5a. We obtained the QM/MM geometries for the reactant (R), transition state (TS) and product (P) that were computed at the MPW1K level of theory using a combination of the 6-31G** and 6-31+G* basis sets,[14] as well as the OPLS force field for the classical region.[16] Single point and geometry optimization QM/MM calculations of these structures were then performed using various SE methods (Table C8) contained within the CHARMM/MNDO97[17] interface. The MM was treated with the OPLS force field[16] in order to ensure that the calculations are in line with those performed by Tvaroška et al.[14] The energy barriers for the reaction are provided in Figure 7.5b. The energies obtained for the MPW1K level are 19.6 kcal/mol (TS) and -5.8 kcal/mol (P) relative to R. The QM region utilized here could not be as large as that defined by Tvaroška et al.[14] (198 atoms), due to MNDO97 limitations. Therefore we made use of an 81 atom QM region (UDG-GlcNAc, Ser21 and His498). The barriers obtained from the QM/MM calculations are provided in Figure 7.5b. AM1*-CB1 underestimates the TS barrier producing and energy of 12.8 kcal/mol, while it overestimates the P barrier giving an energy of -5.4 kcal/mol. This does, however, give small errors of -6.8 and 0.4 kcal/mol for the TS and P, respectively, when compared to the DFT (MPW1K) available data.

Single point calculations were also performed, but not presented here (Table C8, Appendix C).

7.3 Conclusion

A re-parameterization of the exisiting AM1* Hamiltonian has been carried out using a genetic algorithm. The new method, entitled AM1*-CB1, follows the same theory as AM1* and makes use of the H, C, N, and O parameters of AM1/d-CB1. AM1*-CB1 uses a diatomic based core-core repulsion term for the description of phosphorus and as such has a set of new

parameters for this atom and its corresponding interactions, which are significant to chemical glycobiology. The properties which have been parameterized during the development of this method include heats of formation, proton affinities, dipole moments, ionization potential, and interaction energies.

The method is not as accurate as AM1/d-PhoT or AM1/d-CB1 when it comes to predicting proton affinities of phosphate systems and has mixed success with the proton affinities of carbohydrate-phosphate molecules. For dipole moments AM1*-CB1 surpasses all of the methods considered here, for all carbohydrate-phosphate systems, while ionization potentials are lacks accurate. For the heats of formation the method is more accurate than AM1/d-PhoT for all phosphate systems, while it once again has mixed success with the carbohydrate-phosphate systems.

Despite the methods performance for the various properties considered in this work, the method does produce low barrier heights for a reaction relevant to chemical glycobiology with energy barriers of 12.8 and -5.4 kcal/mol for the TS and P, respectively.. It is clear that this method requires additional work in order to be applicable to chemical glycobiological reactions. We believe that one way of getting the method functioning correctly would be to redefine the theory of the method and re-parameterize the phosphorus from thereon. This is a task that shall be considered in the future as it lies outside the scope of this thesis.

7.4 References

(1) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
(2) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. *J. Mol. Model.* **2003**, *9*, 408.
(3) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391.
(4) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1996**, *93*, 315.
(5) Alexeev, Y.; Windus, T. L.; Zhan, C.-G.; Dixon, D. A. *Int. J. Quant. Chem.* **2005**, *102*, 775.
(6) Almerindo, G. I.; Tondo, D. W.; Pliego, J. R. *J. Phys. Chem. A* **2003**, *108*, 166.
(7) Fu, Y.; Liu, L.; Li, R.-Q.; Liu, R.; Guo, Q.-X. *J. Amer. Chem. Soc.* **2003**, *126*, 814.
(8) Hudáky, P.; Perczel, A. *J. Phys. Chem. A* **2004**, *108*, 6195.
(9) Magill, A. M.; Cavell, K. J.; Yates, B. F. *J. Amer. Chem. Soc.* **2004**, *126*, 8717.
(10) Moser, A.; Range, K.; York, D. M. *J. Phys. Chem. B* **2010**, *114*, 13911.
(11) Range, K.; López, C. S.; Moser, A.; York, D. M. *J. Phys. Chem. A* **2005**, *110*, 791.
(12) Range, K.; McGrath, M. J.; Lopez, X.; York, D. M. *J. Amer. Chem. Soc.* **2004**, *126*, 1654.

(13) Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.
(14) Tvaroška, I.; Kozmon, S.; Wimmerová, M.; Koča, J. *J. Am. Chem. Soc.* **2012**, *134*, 15563.
(15) Tvaroška, I.; Kozmon, S.; Wimmerová, M.; Koča, J. *J. Amer. Chem. Soc.* **2012**, *134*, 15563.
(16) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc* **1999**, *121*, 4827.
(17) Thiel, W. University of Zurich, Zurich, Switzerland, 1998.

8. Purine Nucleoside Phosphorylase

Reaction based QM/MM MD simulations are conducted with AM1/d-CB1 to evaluate the performance of the methods when treating a phosphoric based glycobiological system.

8.1 Introduction

The homotrimeric enzyme, purine nucleoside phosphorylase (PNP), catalyzes the reversible phosphorolysis of *β*-nucleosides to free purine and ribose-α-1-phosphate.[1] Although the formation of the nucleoside is usually thermodynamically favored (Figure 8.1), the phosphorolysis direction is favored when the PNP reaction is coupled to purine base oxidation or phosporibosylation (by xanthine oxidase or hypoxanthineguaninephosphoribosyltranferase, respectively) due to rapid metabolic removal of purines. A generic mechanism for PNP[2] is provided in Figure 8.1. It is interesting that uncertainty persists with regards to the extent to which the catalytic process of phosphorolysis resembles the acid-catalyzed hydrolysis; that is, the possible protonation of the imidazole ring N7 by asparagine is not fully documented and alternative mechanisms have been proposed.[3-5]

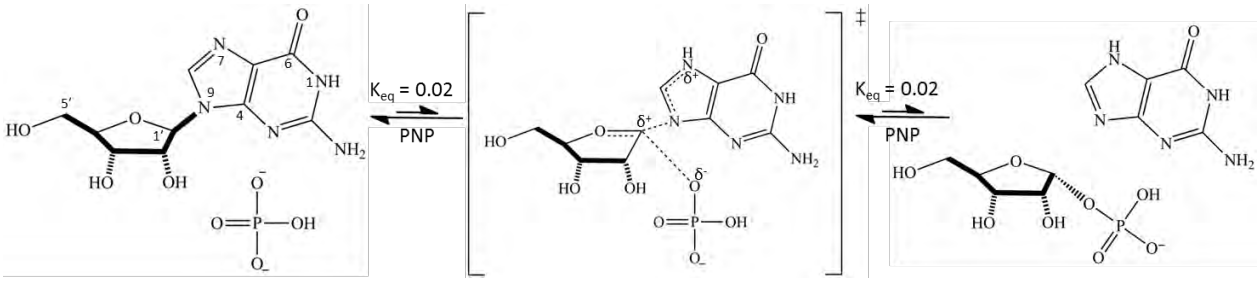


Figure 8.1: Generic mechanism for PNP.

A deficiency in PNP[6] reduces the immune effect of T-cells causing developmental disorders and autoimmune disease. Deoxyguanosine accumulates in the blood as a result of PNP deficiency, and is transported and phosphorylated by T-cell deoxynucleoside kinases to form pathologically elevated levels of deoxy-guanosine-triphosphate (dGTP)[7] specifically in the lymph, causing T-cell apoptosis.[8] T-cells that are over-active can cause certain autoimmune disorders (rheumatoid arthritis, psoriasis, inflammatory bowel disorders, and multiple sclerosis),

tissue transplant rejection and several cancers. Inhibition of PNP can be used to induce T-cell apoptosis, thus PNP has been targeted for rational drug design.[1,2]

Bovine PNP (used in this work) is a homotrimer with P213 symmetry and the active site is located at the interface between subunits (Figure 8.2). Mammalian PNP shares this symmetry, but it is important to point out that not all PNPs do.[9] The secondary structure of PNP is provided in Figure 8.2 along with one of the active sites. It is apparent that the active site would be very exposed if it were not capped by the adjacent monomer.

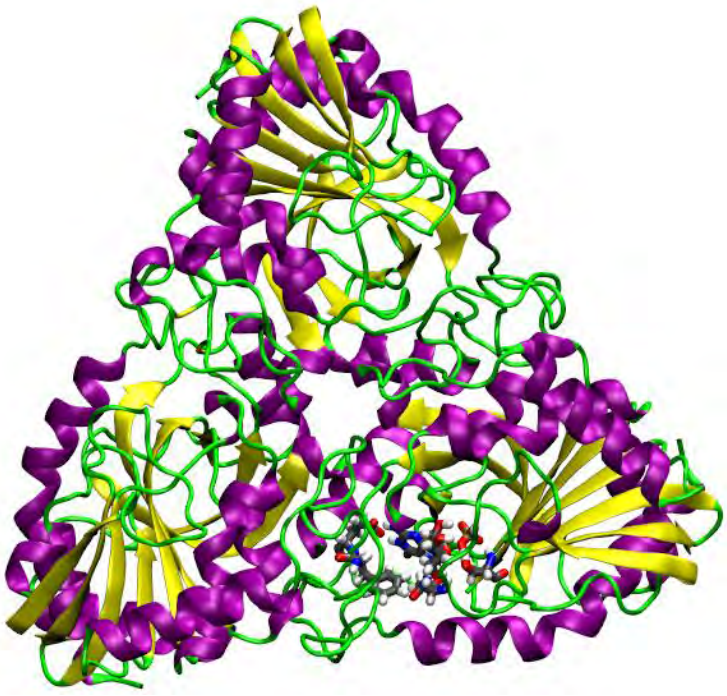


Figure 8.2: Secondary structure of trimeric 1A9S[10] PNP with α-helices in purple, β-sheets in yellow and random coil in green. The molecular structure provided within one of the monomers (represented in licorice) indicates the position of one of the active sites.

Inhibitors of PNP were first developed by using a structure based inhibitor design focused on iterative group alignment established from the PNP crystal structures.[11,12] These inhibitors achieved only nanomolar dissociation constants, which limited their effectiveness because greater than 95% continuous inhibition of PNP is required for significant reduction in T-cell function.[8] Another approach for designing enzyme inhibitors is based on the identification of the transition state (TS) structure stabilized by the target enzyme. TS analogues preferentially

bind their cognate enzyme with high affinity. A first-generation TS analogue (immucillin) proved effective for bovine PNP at concentrations between 36 and 71 pM; however, the effectiveness is less pronounced for human PNP.[1,13] A second generation of immucillin, DADMe immulcillin with an extended linkage is more potent than the first-generation TS analogues requiring less than 6 pM for activity. A third generation of inhibitors has been designed that contains an acyclic iminoalcohol to replace the cyclic mimic of ribooxocarbenium ion at the transition states of PNPs. The best third-generation inhibitor is equivalent to the best inhibitors found in previous generations TS analogues.[1,14] Clearly the aim of PNP studies are centered around finding inhibitory drug targets in order to combat T-cell mediated autoimmune diseases.[15]

As mentioned above the most effective inhibitors are typically based on TS analogues. In order to design such inhibitors both the catalytic activity of amino acids and their mechanistic role must be determined. Erion et al.[2] studied the mechanism by implementing a model of the active site that included only amino acids that interacted closely with the substrate. Using this model a number of active site mutants were created, with the Asn243Ala mutant resulting in a 1000-fold decrease in the k_{cat} for inosine phosphorolysis. This result together with the crystallographic location of the Asn243 side chain suggested a potential TS structure involving hydrogen bond donation by the carboxamido group of Asn243 to N7 of the purine base. Figure 8.3 illustrates Erion's site model, with the substrate in the active site and the surrounding amino acids.

Apart from experimentally determined data, such as kinetic isotope studies,[16-19] a theoretical understanding of the TS is vital for designing active site inhibitors. Barnett and Naidoo[1] have successful run simulations of bovine PNP in which they treated only the nucleoside, present in the active site, quantum mechanically (SCC-DFTB). The authors were unable to model the complete reaction (Figure 8.1) due to the lack of appropriate parameters for SCC-DFTB. In this chapter we look to try and simulate the PNP reaction using the newly parameterized AM1/d-CB1.

Hybrid QM/MM MD simulations, in which the QM region consisted of a number of molecules present in the active site, were conducted with the aid of FEARCF.[20,21]

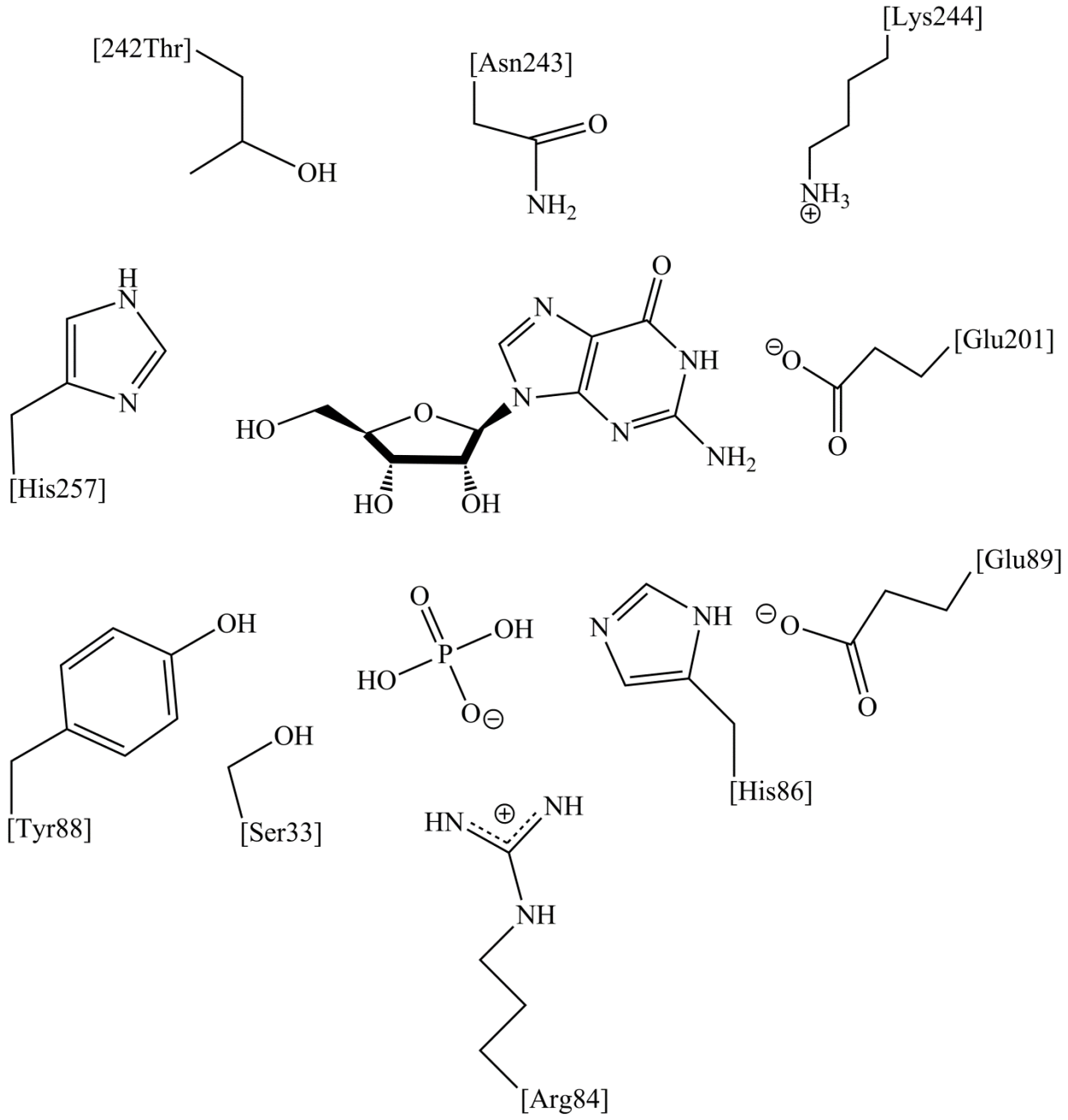


Figure 8.3: PNP active site model constructed by Erion et al.[2] from the atomic coordinates for the PNP-guanine complex (PNP4).

8.2 Simulation details

The 1A9S structure of Ealick,[10] with resolution of 2.00Å, was protonated following a pK_a analysis. Several amino acids are conserved across PNP from different organisms.[4] The trimeric

form was built using the SYMMETRY records in the PDB and the atoms were placed using VMD.[22] A particular active site of the three available was chosen. The waters of crystallization were not removed.[23] $HPO4^{2-}$ was modelled in the binding pocket. Guanosine, Phe200, Glu201, Asn243, Ser33, His86 and $HPO4^{2-}$ made up the QM region of the active site (Figure 8.4). The GHO[24] method was used to join the QM and MM regions. The amino acids selected were based on mutation studies, their conservation across species and on the ability to interact with sugar, base and phosphate parts of the active site.

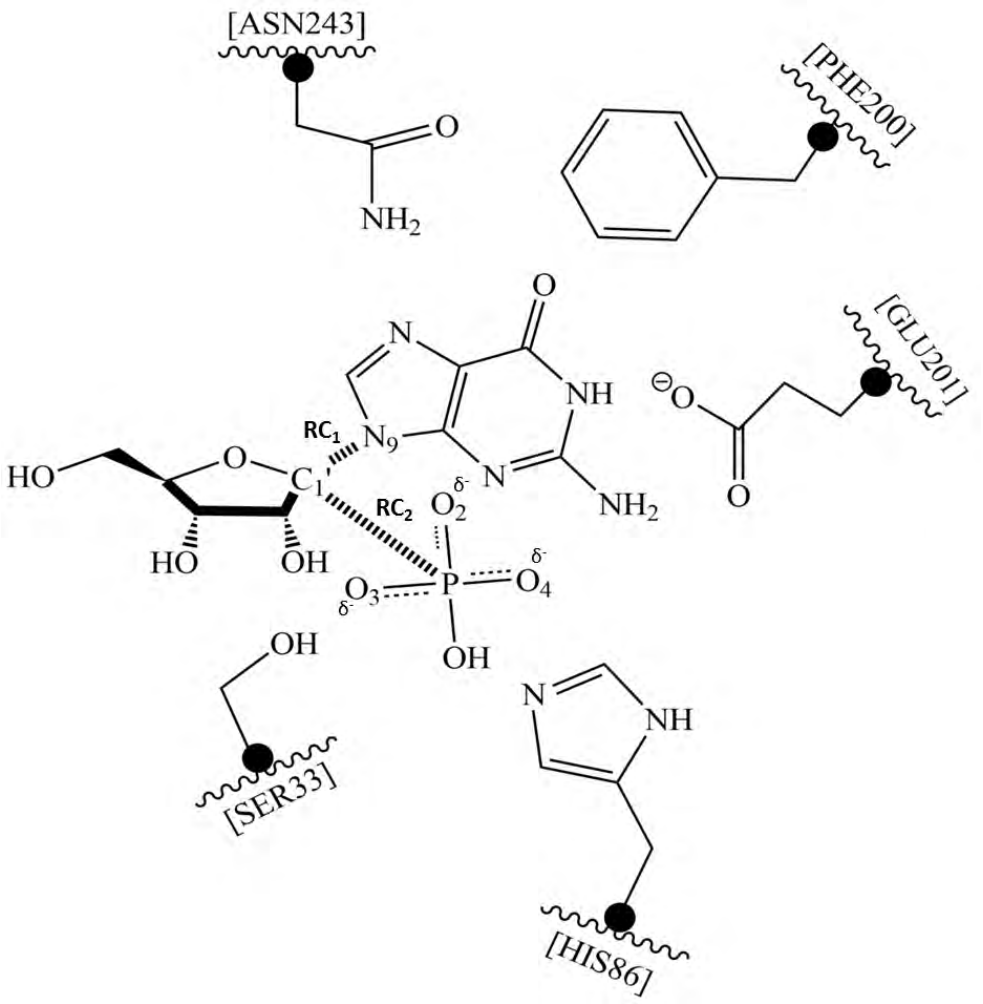


Figure 8.4: Chosen active site for PNP with QM region indicated. GHO atoms are represented with black spheres. RC1 and RC_2 indicate the two reaction coordinates used for the reaction.

After an initial minimization, with the aid of the CHARMM/MNDO97[25,26] interface, a 24.5 Å TIP3P water sphere was positioned over the chosen active site. After heating and equilibration, it was discovered that there was one TIP3P water molecule present in close proximity to the active site given above. As such it was decided to include this water molecule into the QM region together with the active site. An additional equilibration step was conducted after including the water. The equilibrated coordinate and protein structure files were employed to start the free energy calculations. Forty eight 30 ps QM/MM protein FEARCF calculations were run for each iteration of the FEARCF reaction simulation. Leap frog langevin dynamics were carried out for all simulations.

One will notice that the two reaction coordinates chosen for this reaction are the C1−N9 bond and C1−P bond. The reason for selecting the C1−P bond instead of C1 with either O2, O3 or O4 is merely to permit the phosphate to freely rotate and establish the best orientation for either of the oxygen's to attack the anomeric carbon.

8.3 Results and discussion

The 2D reaction energy surfaces obtained for AM1/d-CB1 are provided in Figure 8.5. From this one can see that the bond length of reaction coordinate 2 stretches to a distance of 5.5 Å. Upon reaching this distance the C1'-P bond length is then reduced, crossing the transition state barrier at an energy of 24.9 kcal/mol. Unfortunately, as it can be seen from Figure 8.5 this reaction has not converged and as a result there is a limited amount of transition state sampled. Despite using phosphorus as one of the reaction coordinates it was discovered that the final product had O4 of the phosphate coordinating to the C1' (anomeric carbon) of the carbohydrate ring. We find that the transition state (black square) exists at a C1'-P bond length of 2.94 Å, while the C1'-N9 bond length is 2.02 Å.

An analysis of selected trajectory data generated from the free energy simulations provides a possible reason for the stretching C1'-P bond. Adjacent to O4 and O2 lies the hydroxyl groups of the furanose ring, specifically the hydroxyl of C3' and C2', respectively. During the course of the free energy simulations O4 and O2 establish strong hydrogen bonds with the hydroxyl groups (O4'H32' and O2'H22'). These bonds are so strong that when the hydroxyl rotates it pulls the entire phosphate group along with it. This then produces elongated

C1'-P bond length. The only time the hydrogen bond weakens is at an energy of 24.9 kcal/mol (transition state barrier).

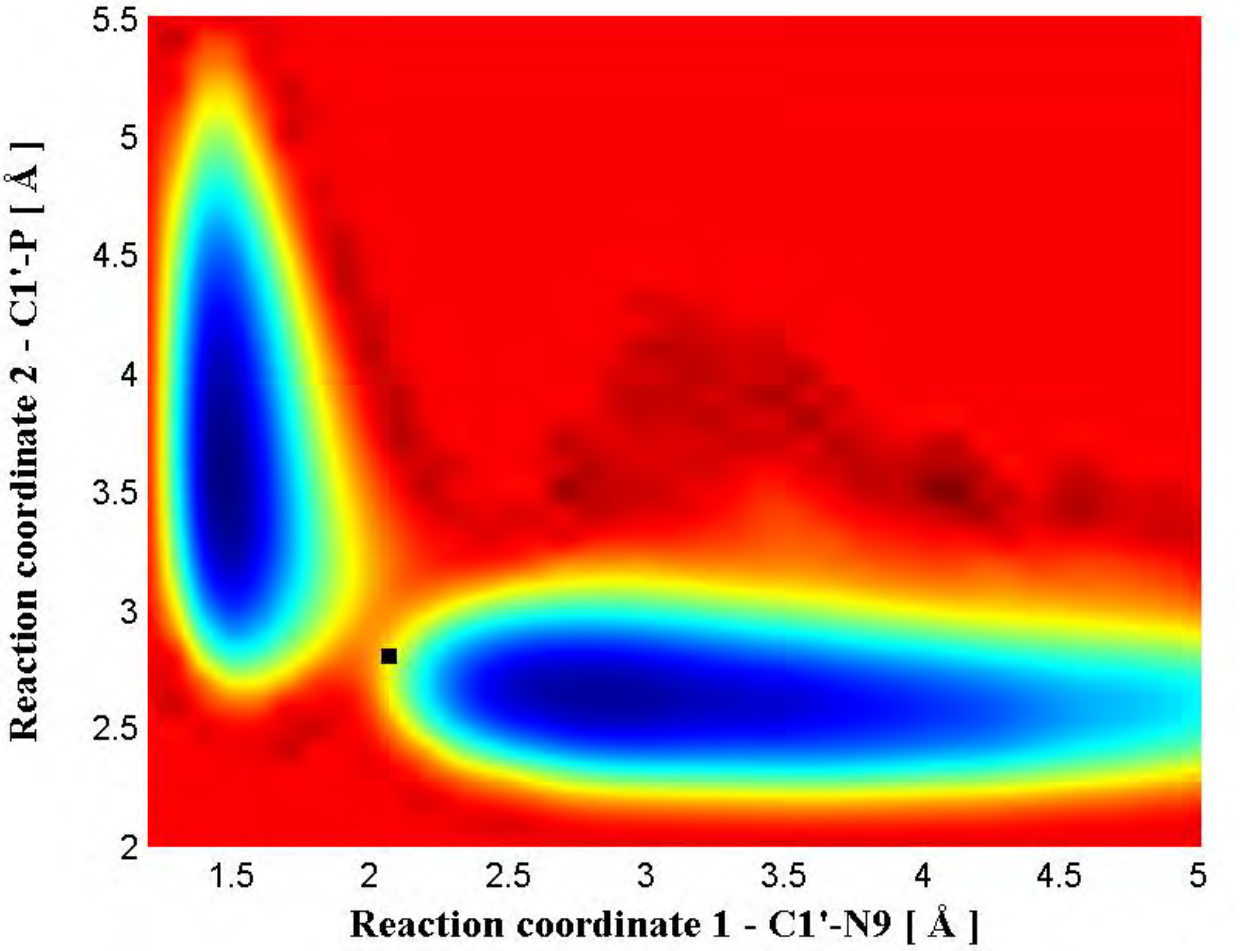


Figure 8.5: AM1/d-CB1 free energy surface viewed along the two reaction coordinates with the transition state indicated by a black square.

The free rotation of the phosphate group was analyzed by monitoring the bond lengths of O2, O3 and O4 with the anomeric carbon (C1'). Figure 8.6 illustrates the variation in bond lengths extracted from various trajectories generated during the free energy simulations. From this we find that each of these oxygen's establish a bond with the anomeric carbon, but the longest lived O-C1' bond is that between O4 and C1'. In fact, as we will show, the final product obtained from the FEARCF simulations has O4 coordinating to C1'.

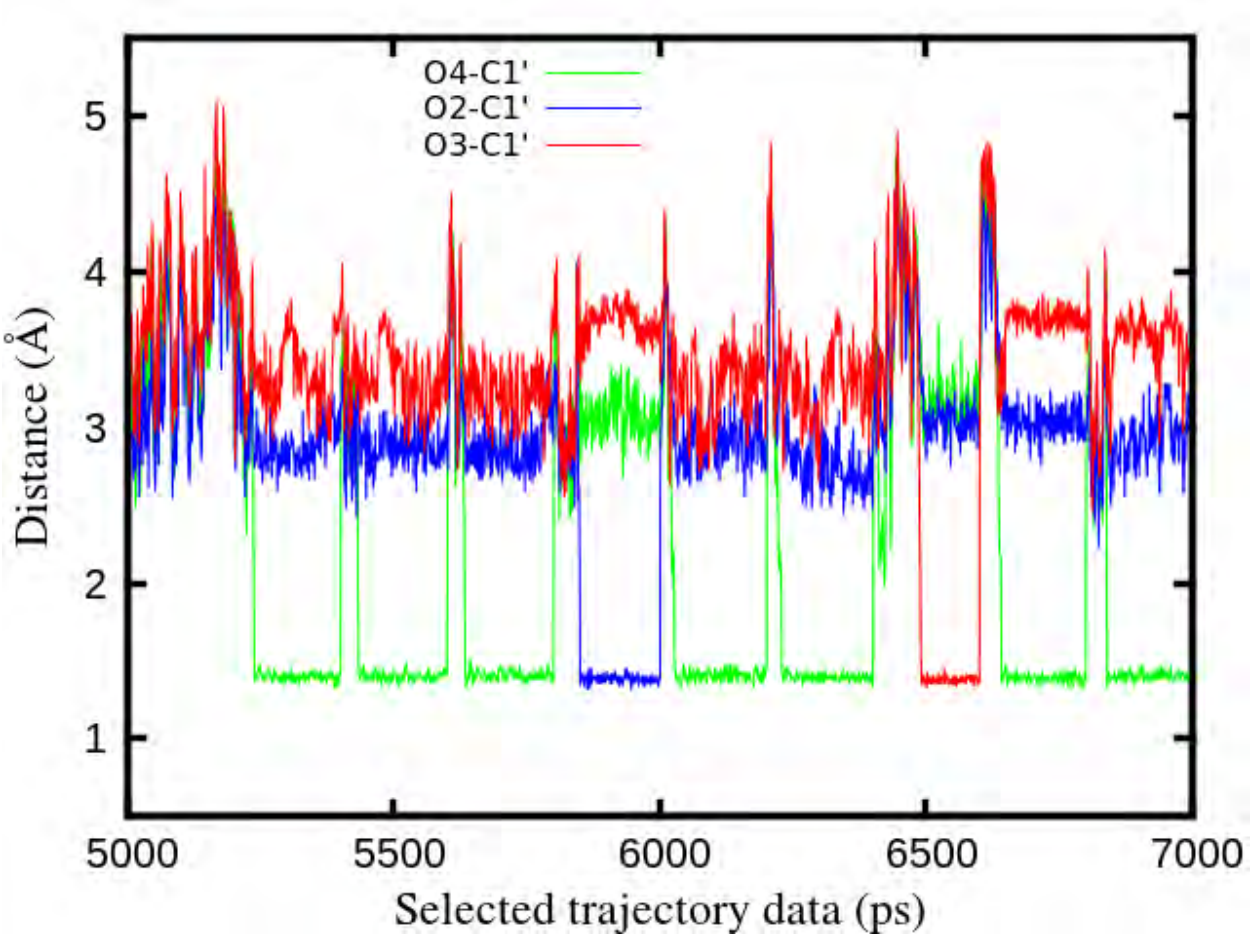


Figure 8.6: Distances of selected FEARCF trajectories for O2-C1', O3-C1' and O4-C1'.

As a further check to determine the product formed in this reaction a number of relevant bond distances were extracted from selected frames of the free energy simulations. The bond lengths focused on specifically were the two reaction coordinates (P-C1' and C1'-N9) as well as the newly formed bond found in the product (O4-C1'). Figure 8.7 provides the reaction coordinates as well as the O4-C1' bond for selected trajectories obtained from FEARCF using AM1/d-CB1. What one will see is that the lifetime of the TS is very short; this is the reason why experimental determination of such a structure is often difficult. However, with the aid of the theoretical techniques utilized in this work we were able to locate a transition state using the newly developed AM1/d-CB1.

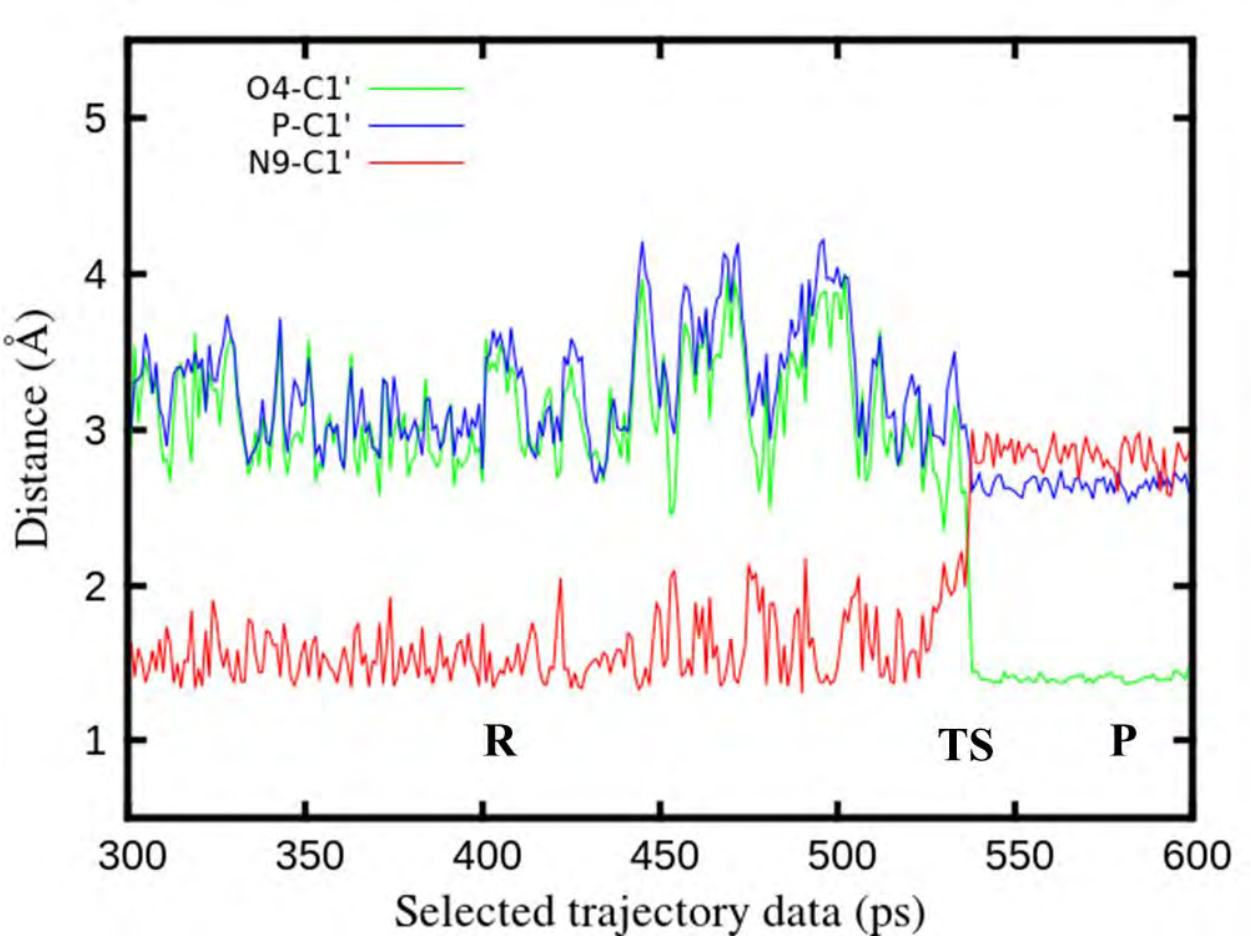


Figure 8.7: The reaction coordinates plotted for selected FEARCF trajectories for AM1/d-CB1 moving from reactant to product.

The TS structure that was acquired from the free energy simulations is provided in Figure 8.8. The bond lengths obtained for the TS are 2.92, 2.02, and 1.44 Å, for the P-C1', N9-C1', and O4-C1' bonds, respectively. Based on the reaction mechanism given in Figure 8.1 it can be seen that the nitrogen (N7) of the base is protonated in the transition state. Due to this protonation state the simulations run in this work had the N7 protonated from the reactant state. The source of the proton is currently unknown, but we suspect it to come from a neighboring water molecule. However, further work will need to be done in order to validate this. Such a reaction will require a 3D FEARCF simulation in which the third reaction coordinate would be N7 with any proton that is in the neighboring vicinity. In this way we ensure that the reactant has N7 deprotonated, while the TS and product will have the species protonated (Figure 8.1).

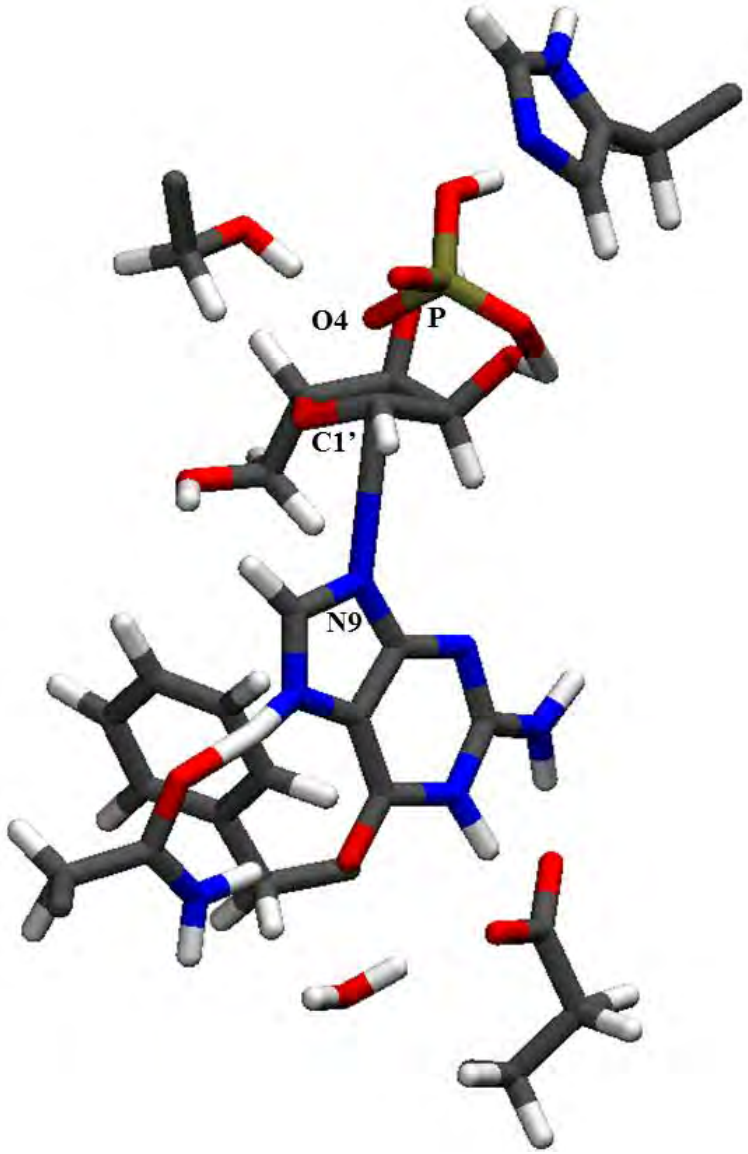


Figure 8.8: Transition state structures obtained from FEARCF simulations of PNP using AM1/d-CB1.

As far as carbohydrate ring puckering is concerned, the TS for AM1/d-CB1 (Figure 8.8) shows the $^{1}T_{2}$ conformation as being the most favorable (definition of 5-membered ring pucker can be found in Chapters 1 and 6). A planar conformer has previously been proposed for this reaction.[27] However, due to the reactions complexity we cannot be certain as to what puckering the ring needs to adhere to in order for this reaction to take place. Further work would need to be

done, using higher levels of theory, in order to validate the most favored conformation for the ring in the TS.

8.4 Conclusion

FEARCF simulations have been conducted on the phosphorylation reaction of bovine PNP using the newly developed AM1/d-CB1 SE methods. By including phosphorus as one of the reaction coordinates we allowed for free rotation around the phosphorus, thereby permitting the most appropriate oxygen to bind to the carbohydrate ring. AM1/d-CB1 has proven its ability to model this reaction with a reasonable transition state barrier of 24.9 kcal/mol. This is, however, a little high in the context of enzymatic reactions. Due to the complexity of PNP there are a number of factors that could constitute this energy barrier: i) Extensive hydrogen bonding between the oxygen of the phosphate and the hydroxyl of the carbohydrate ring. ii) The protonation state of N7 in the reactant (protonated instead of deprotonated as in Figure 8.1). iii) Lack of corrections that allow for more accurate description of hydrogen bonding and dispersion interactions.

To try and improve upon the barrier produced by AM1/d-CB1 one could conduct a 3D reaction energy surface starting N7 of the base as deprotonated and protoning the species as the reaction progresses (third reaction coordinate of 3D FEARCF simulation). In addition appropriate hydrogen bond and dispersion based corrections can be implemented. Such corrections are outside the scope of this thesis, but shall be addressed in future work.

8.5 References

(1) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2013**, *117*, 6019.
(2) Erion, M. D.; Takabayashi, K.; Smith, H. B.; Kessi, J.; Wagner, J.; Hönger, S.; Shames, S. L.; Ealick, S. E. *Biochem.* **1997**, *1997*, 11725.
(3) Tebbe, J.; Bzowska, A.; Wielgus-Kutrowska, B.; Schröder, W.; Kazimierczuk, Z.; Shugar, D.; Saenger, W.; Koellner, G. *J. Mol. Biol.* **1999**, *294*, 1239.
(4) Erion M. D.; Stoeckler J. D.; Guida W. C.; Walter R. L.; Ealick S. E. *Biochem.* **1997**, *36*, 11735.
(5) Canduri, F.; Fadel, V.; Basso, L. A.; Palma, M. S.; Santos, D. S.; de Azevedo Jr, W. F. *Biochem. Biophys. Res. Commun.* **2005**, *327*, 646.
(6) Markert, M. L. *Immunodeficiency Rev.* **1991**, *3*, 45.
(7) Schramm, V. L. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2002**, *1587*, 107.

(8) Kicska, G. A.; Schramm, V. L.; Long, L.; Fairchild, C.; Tyler, P. C.; Furneaux, R. H.; Hãrig, H.; Kaufman, H. L. *Proc. Nat. Acad. Sci.* **2001**, *98*, 4593.
(9) Castilho, M. S.; Postigo, M. P.; Pereira, H. M.; Oliva, G.; Andricopulo, A. D. *Bioorg. Med. Chem.* **2010**, *18*, 1421.
(10) Mao, C.; Cook, W. J.; Zhou, M.; Federov, A. A.; Almo, S. C.; Ealick, S. E. *Biochem.* **1998**, *37*, 7135.
(11) Ealick S. E.; Rule S. A.; Carter D. C.; Greenhough T. J.; Babu, Y. S.; Cook W. J.; Habash J.; Helliwell J. R.; Stoeckler J. D.; Parks R. E. Jr. *J. Biol. Chem.* **1990**, *265*, 1812.
(12) Ealick S. E.; Babu Y. S.; Bugg C. E.; Erion M. D.; Guida W. C.; Montgomery J. A.; D., S. J. A. *Proc. Nat. Acad. Sci. USA* **1991**, *88*, 11540.
(13) Lewandowicz, A.; Tyler, P. C.; Evans, G. B.; Furneaux, R. H.; Schramm, V. L. *J. Biol. Chem.* **2003**, *278*, 31465.
(14) Taylor, E. A.; Clinch, K.; Kelly, P. M.; Li, L.; Evans, G. B.; Tyler, P. C.; Schramm, V. L. *J. Amer. Chem. Soc.* **2007**, *129*, 6984.
(15) Barnett, C. B. PhD Thesis, University of Cape Town, 2010.
(16) Rodgers, J.; Femec, R. J.; Schowen, J. *J. Amer. Chem. Soc.* **1982**, *104*, 3263.
(17) Cleland, W. W. *Methods Enzymol.* **1982**, *87*, 625.
(18) Northrop, D. B. *Biochem.* **1975**, *14*, 2644.
(19) Schramm, V. L. *Ann. Rev. Biochem.* **1998**, *14*, 693.
(20) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243.
(21) Strümpfer, J.; Naidoo, K. J. *J. Comput. Chem.* **2010**, *31*, 308.
(22) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.
(23) Saen-oona, S.; Quaytman-Machledera, S.; Schramm, V. L.; Schwartz, S. D. *Proc. Nat. Acad. Sci.* **2008**, *105*, 16543.
(24) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.
(25) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.
(26) Thiel, W. *MNDO97*, version 5.0; University of Zurich, Zurich, Switzerland, 1998.
(27) Shi, W.; Ting, L.-M.; Kicska, G. A.; Lewandowicz, A.; Tyler, P. C.; Evans, G. B.; Furneaux, R. H.; Kim, K.; Almo, S. C.; Schramm, V. L. *J. Biol. Chem.* **2004**, *279*, 18103.

9. Conclusion

Some concluding remarks related to the methods developed and their application to chemical glycobiology. Future work and improvements that shall be done to these methods is also mentioned.

Two new semi-empirical (SE) methods have been introduced in this thesis, AM1/d-CB1 and AM1*-CB1, which are aimed at modeling systems relevant to chemical glycobiology. The methods make use of a standard *sp* basis and identical parameters when treating pure organic systems. Treatment of hypervalent systems, such as phosphates, involve using an *spd* basis. For AM1/d-CB1 and AM1*-CB1 a different set of phosphorus parameters have been derived due to the slightly different manner in which the two methods treat the core-core repulsion of two atoms (where one of the atoms has *d*-orbitals).

The new methods have been tested within the context of the training set utilized during parameterization. An additional set of testing was performed on various carbohydrate monomers, amino acids, amino acid base pairs, π-stacked sugar-protein interactions. Both methods produce accurate barrier height prediction for QM/MM calculations of a glycosyltransferase reaction when compared to DFT based results. AM1*-CB1 does, however, produce poor results for various properties considered during parameterization.

Due to the poor results obtained for AM1*-CB1 after re-parameterization (properties not as accurate as those produced by other NDDO methods), QM/MM MD simulations of purine nucleoside phosphorylase was only run with AM1/d-CB1. The barrier height achieved for this reaction is 24.9 kcal/mol, which is within range of a typical enzymatic reaction. As such, AM1/d-CB1 shows promise for accurately modelling reactions significant to chemical glycobiology within a QM/MM framework.

9.1 Future work

Even though AM1/d-CB1 produces a reasonable barrier for the PNP reaction, the method does require improvement. It will be beneficial to include empirical based dispersion and hydrogen bond corrections onto the method, together with a re-parameterization of the terms that

make up these corrections. This will ensure that the correction terms work efficiently with the AM1/d-CB1 Hamiltonian.

AM1*-CB1 can be improved upon by redefining the theory of the method as far as treatment of the phosphorus is concerned, thereafter re-parameterizing the method. In addition AM1*-CB1 should also be augmented with corrections for dispersion and hydrogen bonding due the lack thereof in NDDO type methods.

Appendices

Appendix A

Supporting information for results provided in Chapter 5

Appendix B

Supporting information for results provided in Chapter 6

Appendix C

Supporting information for results provided in Chapter 7

Appendix A

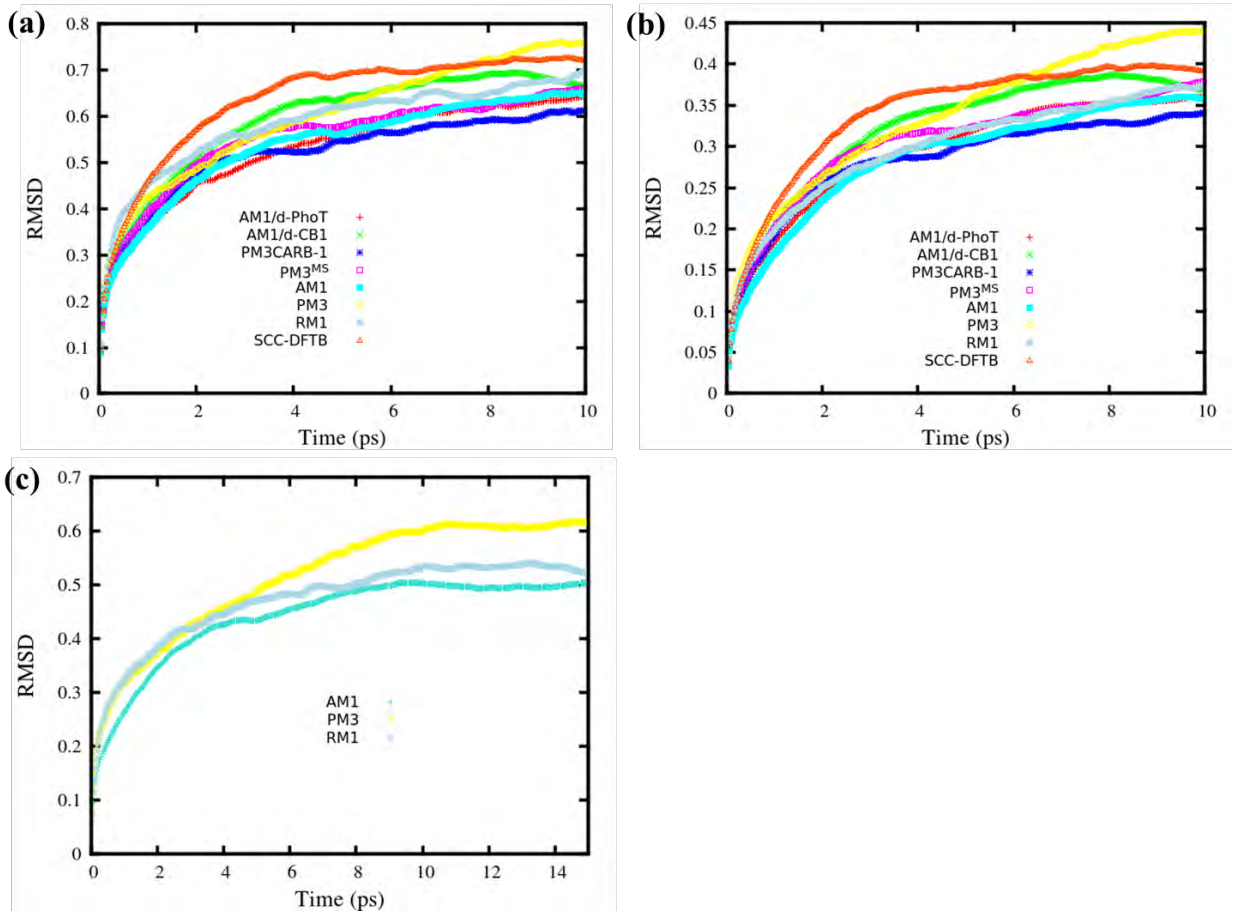


Figure A1. RMSD for atoms (a) C_1 and (b) C_4 of tetrahydrofuran using reference plane of triangular tessellation, as well as (c) average RMSD.

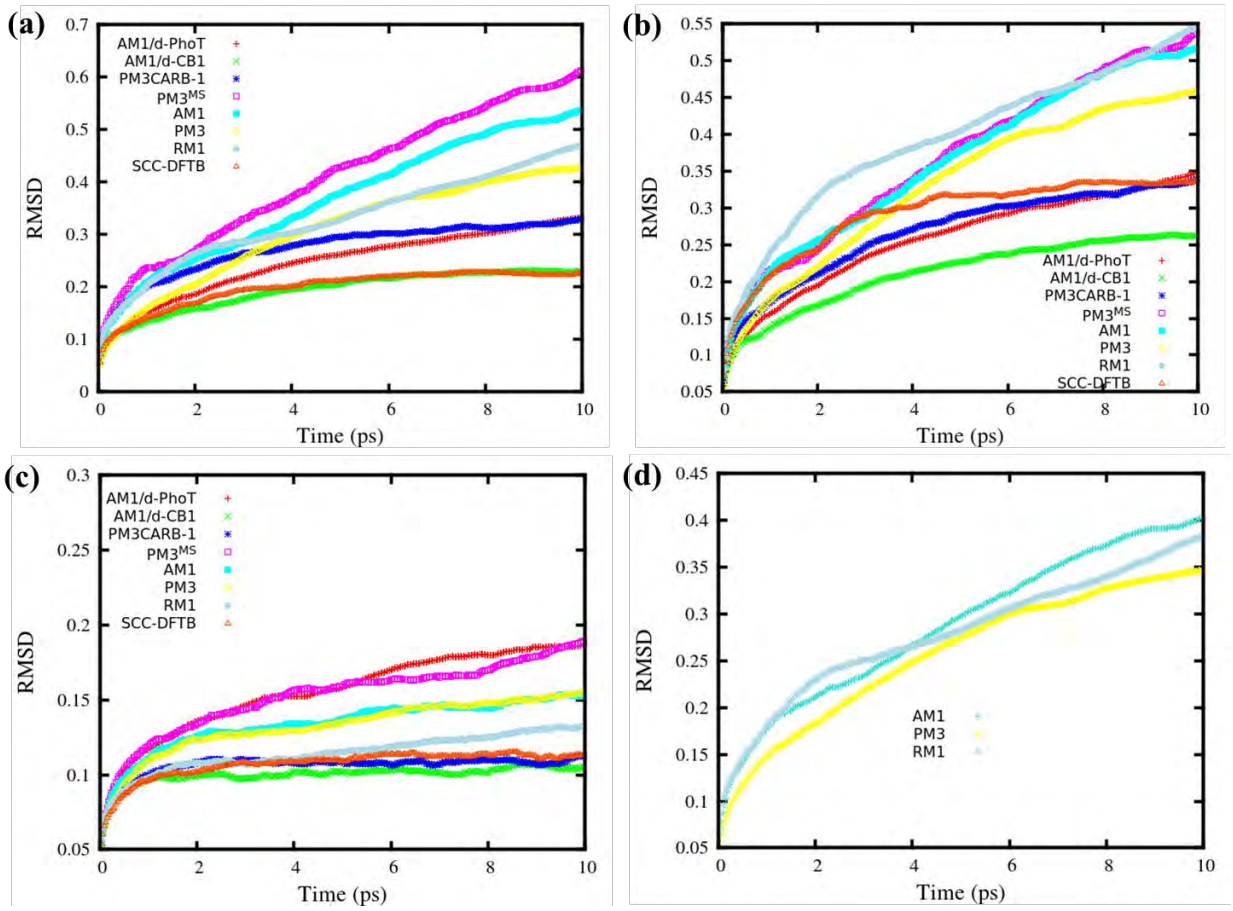


Figure A2. RMSD for atoms (a) C_1, (b) C_3 and (c) C_5 of β-D-glucopyranose using reference plane of triangular tessellation, as well as (c) average RMSD.

Table A1. Original parameters for Hydrogen, Carbon, Nitrogen, Oxygen and Phosphorus derived for the RM1 and AM1/d-PhoT Hamiltonians

Parameters	Hydrogen[a]	Carbon[a]	Nitrogen[a]	Oxygen[a]	Phosphorus[b]
U_{ss}	−11.960677	−51.7255603	−70.8512372	−96.9494807	−46.250810
U_{pp}		−39.4072894	−57.9773092	−77.8909298	−40.712918
U_{dd}					−24.504161
ζ_s	1.08267366	1.85018803	2.37447159	3.17936914	1.909168
ζ_p		1.76830093	1.97812569	2.55361907	2.008466
ζ_d					0.840667
β_s	−5.76544469	−15.4593243	−20.8712455	−29.8510121	−11.194791
β_p		−8.23608638	−16.6717185	−29.1510131	−11.985621
β_d					−2.360095
A	3.06835947	2.79282078	2.96422542	4.17196717	1.883237
G_{ss}	13.98321296	13.0531244	13.08736234	14.00242788	14.645747
G_{pp}		10.95113739	13.69924324	14.14515138	11.694918
G_{sp}		11.33479389	13.21226834	14.95625043	5.689654
G_{p2}		9.72395099	11.94103953	12.70325497	10.328696
H_{sp}		1.55215133	5.00000846	3.93217161	1.175115
ζ_{sn}					2.08512
ζ_{pn}					1.535336
ζ_{dn}					1.236266
ρ_{core}					1.18513
G_{scale}	1.00000000	1.00000000	1.00000000	1.00000000	0.353722
FN_{11}	0.10288875	0.07462271	0.0607338	0.23093552	−0.344529
FN_{21}	5.90172268	5.73921605	4.58892946	5.21828736	3.034933
FN_{31}	1.17501185	1.04396983	1.37873881	0.90363555	1.134275
FN_{12}	0.06457449	0.01177053	0.02438558	0.05859873	−0.021847
FN_{22}	6.41785671	6.92401726	4.62730519	7.42932932	1.684515
FN_{32}	1.93844484	1.66159571	2.08370698	1.5175461	2.716684
FN_{13}	−0.03567387	0.03720662	−0.0228343		−0.036003
FN_{23}	2.80473127	6.26158944	2.05274659		5.243357
FN_{33}	1.63655241	1.63158721	1.86763816		1.924175
FN_{14}		−0.00270657			
FN_{24}		9.00003735			
FN_{34}		2.79557901			

[a] Parameters obtained from the original RM1 Hamiltonian. [b] Parameters obtained from the original AM1/d-PhoT Hamiltonian.

Table A2. Ring relaxation times for molecules used in parameterization (seconds)

	SCC-DFTB[a]	AM1	PM3	RM1
Tetrahydrofuran				
τ	0.47214	0.12944	0.14344	0.23329
β-D-glucopyranose				
τ	0.17384	0.23392	NONE	0.16266

[a] Theoretical values obtained with gas-phase SCC-DFTB[ref 5 of Chapter 5] MD simulations. [b] Relaxation time could not be established since exponential fit was not possible with data generated from the dynamics run. τ corresponds to relaxation time for different pucker angles described in section 4.4.6 of Chapter 4. NONE implies that a relaxation time could not be obtained within the simulation time frame.

Table A3. Molecular classes used during the parameterization

Class	Molecules
Organic	H_2O, H_3O^+, CH_3OH, C_2H_5OH, C_6H_5OH, CH_3CO_2H
Amino acid	Alanine, Arginine, Asparagine, Aspartic acid, Glutamic acid, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Phenylalanine, Proline, Serine, Threonine, Tryptophan, Tyrosine, Valine
Phosphate	HPO_3, HPO_4^{2-}, $H_2PO_4^-$, H_3PO_4, $(OCH_3)(OH)_2PO$, $(OCH_3)(OH)(O)PO^-$, $(OCH_3)_2(OH)PO$
β-gluc−phos	$^1C_4$-HPO_4^-, $^4C_1$-HPO_4^-, $^1C_4$-H_2PO_4, $^4C_1$-H_2PO_4
α-gluc−phos	$^1C_4$-HPO_4^-, $^4C_1$-HPO_4^-, $^1C_4$-H_2PO_4, $^4C_1$-H_2PO_4
β-ribo−phos	6-HPO_4^-, 12-HPO_4^-, 6-H_2PO_4, 12-H_2PO_4
α-ribo−phos	2-HPO_4^-, 4-HPO_4^-, 2-H_2PO_4, 4-H_2PO_4

Table A4. Experimental and Theoretical gas phase proton affinities for molecules used in parameterization (kcal/mol)

	Reference		Error						
Molecule	Exp[a]	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
H_2O	390.3	393.7	20.6	11.4	-73.6	7.2	23.8	4.7	13.6
H_3O^+	165	164.3	-3.8	-12.5	-72.4	3.6	2.1	0.4	3.6
CH_3OH	381.5	382.4	2.2	-1.8	-102.1	-7.5	4.9	2.0	-10.1
C_2H_5OH	378.2	379.2	3.8	-0.4	-95.5	-4.2	5.4	2.4	-7.0
C_6H_5OH	350.1	348.3	-4.0	-7.2	-85.4	-14.9	-1.2	-3.7	-3.7
CH_3CO_2H	347.2	347.2	4.2	0.1	-86.4	-7.8	4.7	-6.4	-4.6
MUE (vs exp)		**1.3**	**6.4**	**5.6**	**85.9**	**7.5**	**7.0**	**3.3**	**7.1**
MSE (vs exp)		**0.5**	**3.8**	**-1.7**	**-85.9**	**-3.9**	**6.6**	**-0.1**	**-1.4**
HPO_3		311.0	18.9	31.4	-48.7	25.0	17.8	-0.5	-3.0
HPO_4^{2-}		584.9	29.1	36.8	-53.7	21.2	27.9	5.2	11.8
$H_2PO_4^-$		460.0	14.9	22.2	-61.8	14.2	11.9	-4.9	0.2
H_3PO_4		329.9	6.1	11.4	-60,3	16.8	7.0	-4.9	-1.2
$(OCH_3)(OH)_2PO$		331.1	4.8	8.9	-65.8	10.8	4.9	-3.4	-4.2
$(OCH_3)(OH)(O)PO^-$		455.3	15.0	20.8	-64.5	12.7	9.9	-1.6	-1.5
$(OCH_3)_2(OH)PO$		330.1	45.7	43.8	117.5	-29.3	43.1	42.5	30.4
MUE (vs DFT)			**19.2**	**25.0**	**67.5**	**18.6**	**17.5**	**9.0**	**7.5**
MSE (vs DFT)			**19.2**	**25.0**	**-33.9**	**10.2**	**17.5**	**4.6**	**4.7**
	Exp[c]	G3MP2[d]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
Alanine ($C_3NH_7O_2$)	215.5	0	-12.2	-13.3	-74.9	-32.3	-0.5	-6.8	-4.0
Arginine ($C_6N_4H_{14}O_2$)	251.2	-1.1	-7.8	-28.8	-87.5	-43.9	-0.8	3.3	-7.3
Asparagine ($C_4N_2H_8O_3$)	222.0	1.6	-3.6	-4.4	-59.0	-19.8	8.8	4.5	7.4
Aspartic acid ($C_4NH_7O_4$)	217.2	1.8	-11.0	-7.9	-70.8	-24.5	0.9	-6.8	-1.6
Glutamic acid ($C_5NH_9O_4$)	218.2	8.3	-10.1	-6.6	-70.5	-20.2	5.1	-0.9	3.9
Glutamine ($C_5N_2H_{10}O_3$)	224.1	8.5	-10.0	-6.6	-67.4	-18.8	8.0	3.9	6.4
Glycine ($C_2NH_5O_2$)	211.9	0	-13.9	-13.6	-74.3	-31.8	-1.8	-9.7	-4.3
Histidine ($C_6N_3H_9O_2$)	236.1	-2.2	-5.7	-11.5	-66.6	-31.9	1.9	-0.1	2.0
Isoleucine ($C_6NH_{13}O_2$)	219.3	0.2	-12.9	-13.2	-76.5	-32.8	-1.3	-6.0	-5.4
Leucine ($C_6NH_{13}O_2$)	218.6	-0.1	-12.2	-13.4	-76.6	-33.0	-0.9	-6.1	-5.1
Lysine ($C_6N_2H_{14}O_2$)	238.0	1.1	-14.5	-17.4	-83.0	-36.6	1.5	-1.4	2.8
Phenylalanine ($C_9NH_{11}O_2$)	220.6	0.6	-11.3	-12.5	-73.9	-33.0	-1.3	-5.2	-4.9
Proline ($C_5NH_9O_2$)	220.0	5.1	-2.1	-12.9	-70.8	-29.7	0.9	2.9	-3.7
Serine ($C_3NH_7O_3$)	218.6	-0.5	-14.1	-15.3	-72.4	-30.1	-3.4	-6.4	-3.1
Threonine ($C_4NH_9O_3$)	220.5	-0.8	-12.2	-13.3	-72.0	-28.5	-1.7	-3.3	-2.2
Tryptophan ($C_{11}N_2H_{12}O_2$)	226.8	-2.1	-4.8	-13.2	-70.0	-32.4	-0.8	0.3	-3.3
Tyrosine ($C_9NH_{11}O_3$)	221.0	1	-11.6	-12.4	-73.5	-31.1	-1.5	-7.1	-5.2
Valine ($C_5NH_{11}O_2$)	217.6	0.9	-11.4	-11.5	-75.0	-30.7	0.3	-4.4	-4.0
MUE (vs. Exp)		**2.0**	**10.1**	**12.7**	**73.0**	**30.1**	**2.3**	**4.4**	**4.3**
MSE (vs. Exp)		**1.2**	**-10.1**	**-12.7**	**-73.0**	**-30.1**	**0.7**	**-2.7**	**-1.8**

[a] Experimental values obtained from Nam et al.[ref 8 of Chapter 5] [b] The DFT proton affinities where computed with M06-2X/6-311++G(3df,2p) level of theory and basis set. [c] Experimental proton affinities obtained from the NIST chemistry webbook.[ref 13 of Chapter 5] [d] G3MP2 proton affinities obtained from Gronert et al.[ref 13 of Chapter 5] All errors are computed as $PA^{calc} - PA^{ref}$.

Table A5. Relative gas phase proton affinities for selected molecules used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose-phosphate								
$^1C_4$−HPO_4^-	0	0.40	0	0	0	0	0	0
$^4C_1$−HPO_4^-	12.17	0	2.25	8.80	3.91	22.30	8.49	25.84
$^1C_4$−H_2PO_4	0	0	0	0	0	0	0	0
$^4C_1$−H_2PO_4	4.58	9.25	20.15	26.38	24.86	14.58	6.11	4.93
MUE (vs DFT)		**4.31**	**6.37**	**6.29**	**7.14**	**5.03**	**0.54**	**3.51**
MSE (vs DFT)		**-1.78**	**1.41**	**4.61**	**3.01**	**5.03**	**-0.54**	**3.51**
β-D-ribofuranose-phosphate[a]								
6−HPO_4^-	3.09	3.45	8.59	13.63	13.28	0.23	0.83	3.95
12−HPO_4^-	0	0	0	0	0	0	0	0
6−H_2PO_4	0.23	2.57	3.20	0.55	1.49	0.98	4.25	2.15
12−H_2PO_4	0	0	0	0	0	0	0	0
MUE (vs DFT)		**0.68**	**2.12**	**2.72**	**2.86**	**0.90**	**1.57**	**0.69**
MSE (vs DFT)		**0.68**	**2.12**	**2.72**	**2.86**	**-0.53**	**0.44**	**0.69**
α-D-glucopyranose-phosphate								
$^1C_4$−HPO_4^-	0	0	0	0	0	0	0	0
$^4C_1$−HPO_4^-	1.56	17.56	15.36	17.40	10.92	6.36	7.08	2.71
$^1C_4$−H_2PO_4	0	1.86	0	0	0	0	0	0
$^4C_1$−H_2PO_4	4.34	0	4.53	10.91	8.69	8.88	0.82	12.84
MUE (vs DFT)		**5.55**	**3.50**	**5.60**	**3.43**	**2.34**	**2.26**	**2.41**
MSE (vs DFT)		**3.38**	**3.50**	**5.60**	**3.43**	**2.34**	**0.50**	**2.41**
α-D-ribofuranose-phosphate								
2−HPO_4^-	5.09	9.57	21.28	26.90	21.13	7.39	0.91	0
4−HPO_4^-	0	0	0	0	0	0	0	0.29
2−H_2PO_4	0.40	0.14	1.88	6.13	4.16	3.86	0.59	0
4−H_2PO_4	0	0	0	0	0	0	0	0.45
MUE (vs DFT)		**1.19**	**4.42**	**6.89**	**4.95**	**1.44**	**1.09**	**1.56**
MSE (vs DFT)		**1.06**	**4.42**	**6.89**	**4.95**	**1.44**	**-1.00**	**-1.19**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT proton affinities obtained with M06-2X/6-311++G(3df,2p).

Table A6. Relative gas phase proton affinities for β-D-carbohydrate conformers used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose								
$^{1}C_{4}$	4.54	0	0	0	0	0	0	0
$^{4}C_{1}$	0	3.50	3.57	0.88	2.84	2.11	6.50	0.10
MUE (vs DFT)		**4.02**	**4.06**	**2.71**	**3.69**	**3.33**	**5.52**	**2.32**
MSE (vs DFT)		**-0.52**	**-0.49**	**-1.83**	**-0.85**	**-1.22**	**0.98**	**-2.22**
^{14}B	10.03	0	7.22	12.08	20.12	8.15	4.04	12.28
^{25}B	6.24	6.69	7.36	6.63	13.14	6.07	1.95	4.78
$^{2}S_{6}$	5.71	5.30	5.10	9.08	10.24	5.76	2.75	8.26
$^{3}S_{1}$	5.62	1.74	2.54	11.85	12.05	4.50	0.23	9.05
^{5}E	5.56	3.97	2.55	5.78	7.47	4.76	0.72	5.35
$^{5}S_{1}$	6.12	12.83	10.31	9.74	12.33	10.16	7.74	6.12
B_{14}	5.54	8.12	7.33	9.44	14.10	7.46	4.27	6.77
$^{2}H_{3}$	0	6.57	3.61	0.83	0	3.92	2.75	0.45
$^{6}H_{1}$	1.14	2.38	0	0	2.43	0	0	0
MUE (vs DFT)		**3.72**	**2.37**	**2.42**	**5.10**	**1.67**	**3.36**	**1.41**
MSE (vs DFT)		**0.18**	**0.01**	**2.16**	**5.10**	**0.54**	**-2.39**	**0.79**
β-D-ribofuranose[a]								
2	1.68	7.48	6.59	4.36	5.03	5.66	7.04	2.78
4	0.99	6.50	5.23	2.63	3.55	4.33	5.93	1.45
5	0	0	1.61	1.49	3.03	0.26	1.26	0.47
6	0.96	0.18	2.75	3.80	4.93	2.16	2.16	2.91
6b	2.74	2.43	3.41	5.22	7.47	2.74	7.15	4.40
6ab	3.34	2.62	3.40	4.65	6.65	2.26	7.05	3.43
7	0.75	1.30	0	0.31	0	0.65	0.99	0.88
8	6.79	6.12	7.03	7.95	5.78	5.12	5.95	5.97
9	6.69	6.24	6.05	6.34	6.93	5.78	7.34	6.29
10	3.11	4.71	5.84	4.60	7.86	5.08	7.79	4.67
11	1.00	3.45	4.28	1.63	5.26	2.63	5.58	1.60
12	2.80	0.01	1.91	3.65	3.83	1.78	0	3.04
13	3.11	2.52	2.51	2.81	2.57	0.94	1.58	1.55
14	3.69	1.59	1.32	3.75	2.51	0	0.97	2.15
15	0.15	1.51	1.31	0	1.79	0.65	2.26	0
MUE (vs DFT)		**1.71**	**1.73**	**1.19**	**2.42**	**1.57**	**2.74**	**0.85**
MSE (vs DFT)		**0.59**	**1.03**	**1.03**	**1.96**	**0.15**	**1.68**	**0.25**

[a] Initial ring conformations obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table A7. Relative gas phase proton affinities for α-D-carbohydrate conformers used in parameterization (kcal/mol).

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	PM3^{MS}	RM1	AM1/d-PhoT	AM1/d-CB1
α-D-glucopyranose								
$^{1}C_{4}$	1.25	1.88	0	0	0	0	0	0
$^{4}C_{1}$	0	0	0.68	2.49	4.43	1.03	1.69	2.27
MUE (vs DFT)		**0.32**	**0.97**	**1.87**	**2.84**	**1.14**	**1.47**	**1.76**
MSE (vs DFT)		**0.32**	**-0.29**	**0.62**	**1.59**	**-0.11**	**0.22**	**0.51**
^{14}B	11.26	3.38	8.04	16.98	20.00	9.72	12.61	14.90
^{25}B	10.19	7.28	6.95	14.06	13.64	6.67	10.88	8.55
$^{2}S_{6}$	7.69	5.37	4.60	10.67	9.92	7.10	7.18	9.90
$^{3}S_{1}$	6.85	6.75	4.40	18.95	12.80	9.34	9.08	14.80
$^{5}S_{1}$	0.27	1.05	0.08	1.85	4.16	1.72	1.81	1.52
B_{14}	0	0	0	0	2.44	0.91	0	0.19
$^{2}H_{3}$	1.32	2.54	0.56	1.88	0	0	4.84	0
$^{6}H_{1}$	3.95	4.90	2.37	5.61	3.82	2.70	8.03	3.70
B_{36}	8.11	8.79	4.03	5.78	3.26	4.71	8.46	5.48
MUE (vs DFT)		**1.87**	**2.07**	**3.42**	**3.67**	**1.83**	**1.59**	**2.34**
MSE (vs DFT)		**-1.06**	**-2.07**	**2.90**	**2.27**	**-0.75**	**1.47**	**1.05**
α-D-ribofuranose[a]								
1	8.63	8.50	5.30	9.70	7.38	6.58	13.21	9.19
3	0	0	0.39	0	0	0	0	0
7	12.92	11.13	1.92	13.67	4.09	8.48	15.37	14.07
8	12.38	7.52	3.07	10.31	5.59	4.85	12.67	9.49
9	6.78	4.41	1.45	7.96	3.57	4.87	6.80	9.30
10	10.43	10.73	7.51	12.36	7.84	9.51	16.34	12.57
11	8.35	8.62	5.45	9.36	5.02	7.74	12.14	9.56
15	4.16	3.32	0.74	5.04	0.31	3.00	5.92	5.07
16	4.62	2.00	0	5.18	1.45	0.73	4.27	5.01
MUE (vs DFT)		**1.46**	**4.80**	**1.05**	**3.67**	**2.50**	**2.13**	**1.31**
MSE (vs DFT)		**-1.34**	**-4.72**	**0.59**	**-3.67**	**-2.50**	**2.05**	**0.67**

[a] Initial ring conformations obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table A8. Absolute dipole moments of selected molecules used in parameterization (Debye)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
H_2O	1.93	-0.08	-0.17	0.77	0.51	-0.07	0.47	0.26
CH_3OH	1.69	-0.07	-0.19	0.52	0.40	-0.04	0.62	0.25
C_2H_5OH	1.60	-0.07	-0.16	0.55	0.39	-0.04	0.58	0.24
C_6H_5OH	1.30	-0.10	-0.16	0.50	0.26	-0.06	0.45	0.24
CH_3CO_2H	1.77	-0.19	-0.15	0.59	0.21	-0.20	0.29	0.06
CH_3OCH_3	1.32	0.11	-0.05	0.41	0.51	0.15	0.90	0.38
MUE (vs DFT)		**0.10**	**0.15**	**0.56**	**0.38**	**0.09**	**0.55**	**0.24**
MSE (vs DFT)		**-0.07**	**-0.15**	**0.56**	**0.38**	**-0.04**	**0.55**	**0.24**
Aspartic acid ($C_4NH_7O_4$)	2.94	-0.16	-0.10	0.30	0.58	-0.19	0.28	-0.06
Asparagine ($C_4N_2H_8O_3$)	4.98	-0.17	-0.14	1.14	1.06	-0.12	0.90	0.25
Glutamic acid ($C_5NH_9O_4$)	2.78	-0.39	-0.13	0.68	0.68	-0.41	-0.01	-0.12
Glutamine ($C_5N_2H_{10}O_3$)	1.95	-0.15	-0.02	0.56	0.73	-0.11	0.19	-0.05
Histidine ($C_6N_3H_9O_2$)	3.42	-0.08	0.28	0.94	0.46	0.05	0	0.33
Arginine ($C_6N_4H_{14}O_2$)	1.63	-0.18	-0.20	-1.24	-0.52	0.05	-0.40	-0.50
Phenylalanine ($C_9NH_{11}O_2$)	2.26	0.08	-0.03	0.73	0.14	0.10	0.57	0.49
Tyrosine ($C_9NH_{11}O_3$)	2.22	0.04	-0.08	0.12	0.03	0.06	0.30	0.15
Tryptophan ($C_{11}N_2H_{12}O_2$)	3.86	0.01	0.19	0.99	1.06	0.03	0.36	0.38
MUE (vs DFT)		**0.14**	**0.13**	**0.74**	**0.58**	**0.12**	**0.33**	**0.26**
MSE (vs DFT)		**-0.11**	**-0.03**	**0.47**	**0.47**	**-0.06**	**0.24**	**0.10**
HPO_3	3.34	-0.33	-0.78	0.45	0.22	-0.63	0.16	-0.05
$P(CH_3)_3$	1.22	0.53	-0.19	-0.54	-0.25	-0.14	0.91	3.21
$(CH_3)_3P(O)$	4.47	-0.28	-0.51	0.61	0.31	-0.42	0.84	1.20
H_3PO_4	0.49	-0.35	-0.41	-0.10	-0.38	-0.39	-0.43	-0.32
$(OCH_3)(OH)_2PO$	0.94	-0.09	-0.42	-0.40	-0.36	-0.17	0.19	0.07
$(OCH_3)_2(OH)PO$	1.14	-0.07	-0.43	-0.45	-0.32	-0.19	0.13	0.06
$(OCH_3)_3PO$	1.05	-0.21	-0.25	-0.10	-0.07	-0.40	-0.61	-0.27
MUE (vs DFT)		**0.27**	**0.43**	**0.38**	**0.27**	**0.33**	**0.47**	**0.74**
MSE (vs DFT)		**-0.11**	**-0.43**	**-0.08**	**-0.12**	**-0.33**	**0.17**	**0.56**
β-D-glucopyranose-phosphate								
$^1C_4-H_2PO_3$	10.0	-0.67	-1.94	1.54	1.06	-1.01	1.35	0.79
$^4C_1-H_2PO_3$	4.85	0.07	-1.33	0.04	-0.17	-0.38	0.74	0.54
α-D-glucopyranose-phosphate								
$^1C_4-H_2PO_3$	5.16	-0.04	-1.08	0.68	0.41	-0.27	0.94	0.68
$^4C_1-H_2PO_3$	3.73	-0.01	-0.93	0.46	0.05	-0.21	0.76	0.74
MUE (vs DFT)		**0.20**	**1.32**	**0.68**	**0.42**	**0.47**	**0.95**	**0.69**
MSE (vs DFT)		**-0.16**	**-1.32**	**0.68**	**0.34**	**-0.47**	**0.95**	**0.69**
β-D-ribofuranose-phosphate[a]								
$6-H_2PO_3$	3.97	0.05	-1.05	0.04	-0.12	-0.32	0.44	0.44
$12-H_2PO_3$	3.76	0.01	-1.05	0.07	-0.23	-0.38	0.24	0.36
α-D-ribofuranose-phosphate								
$2-H_2PO_3$	6.52	0.03	-1.33	1.01	0.58	-0.35	1.56	0.91
$4-H_2PO_3$	6.64	-0.13	-1.82	0.36	-0.08	-0.63	0.99	0.58
MUE (vs DFT)		**0.06**	**1.31**	**0.37**	**0.25**	**0.42**	**0.81**	**0.57**
MSE (vs DFT)		**-0.01**	**-1.31**	**0.37**	**0.04**	**-0.42**	**0.81**	**0.57**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT dipole moments obtained with M06-2X/6-311++G(3df,2p). All error are computed as $\mu^{calc} - \mu^{DFT}$.

Table A9. Absolute dipole moments of *β*-D-carbohydrate conformers used in parameterization (Debye)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose								
$^{1}C_{4}$	4.20	-0.70	-0.50	1.34	0.90	-0.52	0.55	0.27
$^{4}C_{1}$	2.87	-0.58	-0.33	0.95	0.58	-0.47	0.08	0.12
MUE (vs DFT)		**0.64**	**0.42**	**1.15**	**0.74**	**0.50**	**0.32**	**0.19**
MSE (vs DFT)		**-0.64**	**-0.42**	**1.15**	**0.74**	**-0.50**	**0.32**	**0.19**
^{14}B	3.48	-0.41	-0.41	1.18	0.83	-0.34	0.64	0.28
^{25}B	2.62	-0.48	-0.35	0.81	0.46	-0.39	0.21	0.16
$^{2}S_{6}$	3.49	-0.53	-0.36	1.16	0.82	-0.44	0.42	0.18
$^{3}S_{1}$	2.21	-0.44	-0.28	0.63	0.42	-0.36	0.05	0.06
^{5}E	0.89	0.10	-0.11	0.27	0.27	0.12	0.68	0.26
$^{5}S_{1}$	1.67	-0.35	-0.29	0.48	0.21	-0.31	0.05	0.02
B_{14}	2.19	-0.30	-0.23	0.77	0.40	-0.26	0.23	0.23
$^{1}S_{3}$	1.01	-0.07	-0.02	0.40	0.33	-0.06	0.18	0.09
$^{2}H_{3}$	3.96	-0.60	-0.56	1.12	0.88	-0.46	0.72	0.20
$^{6}H_{1}$	3.41	-0.40	-0.46	1.05	0.68	-0.32	0.69	0.26
B_{36}	1.15	-0.23	-0.22	0.49	0.02	-0.21	0.06	0.14
MUE (vs DFT)		**0.36**	**0.30**	**0.76**	**0.48**	**0.30**	**0.36**	**0.17**
MSE (vs DFT)		**-0.34**	**-0.30**	**0.76**	**0.48**	**-0.28**	**0.36**	**0.17**
β-D-ribofuranose[a]								
1	3.51	-0.61	-0.54	1.00	0.62	-0.48	0.46	0.19
2	2.35	-0.23	-0.25	0.74	0.59	-0.17	0.57	0.25
3	3.89	-0.37	-0.41	1.38	0.92	-0.28	0.89	0.43
4	1.71	-0.30	-0.20	0.53	0.40	-0.24	0.14	0.06
5	4.24	-0.50	-0.40	1.56	1.05	-0.41	0.68	0.34
6	1.26	-0.05	-0.04	0.59	0.28	-0.06	0.18	0.21
6b	3.07	-0.27	-0.32	0.98	0.66	-0.20	0.66	0.38
6ab	3.38	-0.37	-0.41	1.03	0.71	-0.27	0.70	0.37
7	3.11	-0.21	-0.34	1.09	0.85	-0.11	0.98	0.44
8	6.09	-0.70	-0.83	1.87	1.25	-0.52	1.44	0.61
9	3.68	-0.43	-0.50	1.18	0.70	-0.33	0.89	0.43
10	4.50	-0.62	-0.61	1.24	0.82	-0.51	0.66	0.28
11	1.06	-0.29	-0.30	-0.03	0.02	-0.23	0.09	-0.12
12	2.06	-0.59	-0.27	0.63	0.35	-0.54	-0.35	-0.15
13	2.79	-0.48	-0.45	0.92	0.50	-0.37	0.45	0.24
14	3.40	-0.54	-0.45	1.14	0.68	-0.44	0.41	0.22
15	3.16	-0.18	-0.26	1.25	1.00	-0.08	1.12	0.48
16	2.30	-0.47	-0.40	0.51	0.34	-0.42	0.11	-0.05
MUE (vs DFT)		**0.40**	**0.39**	**0.98**	**0.65**	**0.31**	**0.60**	**0.29**
MSE (vs DFT)		**-0.40**	**-0.39**	**0.98**	**0.65**	**-0.31**	**0.56**	**0.26**

[a] Conformations obtained from ref 15 of Chapter 5. [b] DFT dipole moments obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $\mu^{calc} - \mu^{DFT}$.

Table A10. Absolute dipole moments of α-D-carbohydrate conformers used in parameterization (Debye)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-D-glucopyranose								
$^{1}C_{4}$	3.25	-0.26	-0.51	0.85	0.65	-0.18	1.05	0.36
$^{4}C_{1}$	5.58	-0.68	-0.52	2.03	1.46	-0.50	1.01	0.55
MUE (vs DFT)		**0.47**	**0.52**	**1.44**	**1.06**	**0.34**	**1.03**	**0.45**
MSE (vs DFT)		**-0.47**	**-0.52**	**1.44**	**1.06**	**-0.34**	**1.03**	**0.45**
^{14}B	2.78	-0.09	-0.22	0.98	0.75	-0.03	0.91	0.49
^{25}B	3.47	-0.15	-0.25	1.42	0.95	-0.03	1.20	0.67
$^{2}S_{6}$	1.40	-0.33	-0.06	0.52	0.31	-0.26	0	0.08
$^{3}S_{1}$	1.58	-0.08	-0.14	0.64	0.33	-0.04	0.56	0.33
$^{5}S_{1}$	1.40	-0.39	-0.15	0.35	0.16	-0.32	-0.19	-0.02
B_{14}	3.44	-0.51	-0.30	1.26	0.65	-0.40	0.36	0.37
$^{1}S_{3}$	3.05	-0.28	-0.25	1.08	0.81	-0.19	0.64	0.38
$^{2}H_{3}$	4.88	-0.25	-0.40	1.76	1.53	-0.10	1.68	0.70
$^{6}H_{1}$	4.46	-0.20	-0.42	1.70	1.21	-0.08	1.56	0.73
B_{36}	1.49	-0.12	-0.21	0.48	0.20	-0.13	0.41	0.09
MUE (vs DFT)		**0.24**	**0.24**	**1.02**	**0.69**	**0.16**	**0.75**	**0.39**
MSE (vs DFT)		**-0.24**	**-0.24**	**1.02**	**0.69**	**-0.16**	**0.71**	**0.38**
α-D-ribofuranose[a]								
1	4.80	-0.35	-0.48	1.76	1.27	-0.21	1.45	0.66
2	2.38	-0.07	-0.35	0.82	0.45	-0.02	1.01	0.47
3	3.05	-0.33	-0.37	1.05	0.76	-0.22	0.83	0.34
4	0.89	-0.26	-0.30	0.11	-0.13	-0.23	0.16	-0.04
5	2.88	-0.64	-0.36	0.99	0.63	-0.53	0.05	0.00
6	3.20	-0.19	-0.32	1.12	0.75	-0.13	0.83	0.45
6b	3.14	-0.19	-0.31	1.11	0.75	-0.13	0.83	0.45
6ab	3.16	-0.25	-0.37	1.08	0.78	-0.16	0.86	0.42
7	4.11	-0.22	-0.49	1.30	0.98	-0.12	1.30	0.56
8	5.11	-0.39	-0.64	1.75	1.22	-0.23	1.66	0.73
9	2.89	-0.18	-0.57	0.83	0.43	-0.10	1.20	0.48
10	4.26	-0.34	-0.46	1.52	1.03	-0.24	1.12	0.56
11	2.74	-0.45	-0.44	0.86	0.54	-0.34	0.51	0.16
12	1.32	0.02	-0.30	0.27	0.23	0.03	0.78	0.19
13	4.43	-0.39	-0.53	1.60	0.98	-0.29	1.15	0.58
14	4.18	-0.48	-0.51	1.43	0.86	-0.40	0.74	0.37
15	3.68	-0.25	-0.35	1.38	1.08	-0.12	1.27	0.55
16	3.20	-0.54	-0.41	0.99	0.62	-0.47	0.27	0.09
MUE (vs DFT)		**0.31**	**0.42**	**1.11**	**0.75**	**0.22**	**0.89**	**0.39**
MSE (vs DFT)		**-0.31**	**-0.42**	**1.11**	**0.74**	**-0.22**	**0.89**	**0.39**

[a] Initial conformations obtained from ref 15 of Chapter 5. [b] DFT dipole moments obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $\mu^{calc} - \mu^{DFT}$.

Table A11. Absolute ionization potentials of selected molecules used in parameterization (eV)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
H_2O	12.78	-0.32	-0.45	3.93	1.67	-0.53	0.01	-0.35
CH_3OH	11.09	0.09	0.09	1.76	0.75	-0.08	-0.07	0.14
C_2H_5OH	10.78	0.14	0.15	1.41	0.57	-0.02	-0.25	0.16
C_6H_5OH	8.73	0.43	0.49	0.88	1.57	0.26	0.26	-0.15
CH_3CO_2H	10.95	0.69	0.51	3.37	2.11	0.55	1.05	0.86
CH_3OCH_3	10.17	0.47	0.52	2.16	0.85	0.29	0.13	0.65
MUE (vs DFT)		**0.36**	**0.37**	**2.25**	**1.25**	**0.29**	**0.30**	**0.39**
MSE (vs DFT)		**0.25**	**0.22**	**2.25**	**1.25**	**0.08**	**0.19**	**0.22**
Aspartic acid ($C_4NH_7O_4$)	9.71	0.60	-0.02	0.51	-0.14	0.35	0.35	0.21
Asparagine ($C_4N_2H_8O_3$)	9.48	0.63	0.01	0.44	-0.14	0.38	0.32	0.22
Glutamic acid ($C_5NH_9O_4$)	9.98	0.55	-0.01	0.57	-0.13	0.31	0.41	0.24
Glutamine ($C_5N_2H_{10}O_3$)	9.64	0.73	0.24	0.74	0.15	0.45	0.45	0.34
Histidine ($C_6N_3H_9O_2$)	8.71	0.63	0.74	1.06	0.82	0.60	0.33	0.15
Argenine ($C_6N_4H_{14}O_2$)	8.90	0.51	0.13	0.52	-0.10	0.49	0.09	0.20
Phenylalanine ($C_9NH_{11}O_2$)	9.09	0.60	0.46	0.88	0.47	0.38	0.26	0
Tyrosine ($C_9NH_{11}O_3$)	8.56	0.69	0.72	1.24	0.97	0.54	0.52	0.28
Tryptophan ($C_{11}N_2H_{12}O_2$)	7.84	0.72	0.61	0.86	1.32	0.55	0.41	0.13
MUE (vs DFT)		**0.63**	**0.33**	**0.76**	**0.47**	**0.45**	**0.35**	**0.20**
MSE (vs DFT)		**0.63**	**0.32**	**0.76**	**0.36**	**0.45**	**0.35**	**0.20**
$P(CH_3)_3$	8.56	0.81	0.28	0.68	0.04	0.43	-0.90	0.36
$(CH_3)_3P(O)$	9.85	0.99	-0.04	1.39	0.95	0.64	0.01	0.76
H_3PO_4	11.63	0.47	-0.65	3.54	1.42	0.27	0.38	-0.35
HPO_3	12.71	-0.38	-1.52	2.70	0.65	-0.72	0.12	0.56
$(OCH_3)(OH)_2PO$	11.29	0.59	-0.43	1.67	0.87	0.38	0.33	0.69
$(OCH_3)_2(OH)PO$	11.06	0.63	-0.32	1.78	0.96	0.40	0.34	0.76
$(OCH_3)_3PO$	10.90	0.63	-0.25	1.88	1.08	0.41	0.32	0.79
MUE (vs DFT)		**0.64**	**0.50**	**1.95**	**0.85**	**0.46**	**0.34**	**0.61**
MSE (vs DFT)		**0.53**	**-0.42**	**1.95**	**0.85**	**0.26**	**0.09**	**0.51**
β-D-glucopyranose-phosphate								
$^1C_4-H_2PO_3$	10.38	0.61	0.27	1.94	1.32	0.64	0.59	1.09
$^4C_1-H_2PO_3$	10.46	0.69	0.29	2.08	1.24	0.71	0.67	1.17
α-D-glucopyranose-phosphate								
$^1C_4-H_2PO_3$	10.01	0.70	0.54	2.19	1.41	0.77	0.72	1.22
$^4C_1-H_2PO_3$	10.54	0.83	0.28	2.01	1.30	0.83	0.72	1.23
MUE (vs DFT)		**0.71**	**0.35**	**2.06**	**1.32**	**0.74**	**0.68**	**1.18**
MSE (vs DFT)		**0.71**	**0.35**	**2.06**	**1.32**	**0.74**	**0.68**	**1.18**
β-D-ribofuranose-phosphate[a]								
$6-H_2PO_3$	10.59	0.69	0.41	2.05	1.49	0.70	0.67	1.02
$12-H_2PO_3$	10.60	0.65	0.35	2.06	1.40	0.64	0.65	1.02
α-D-ribofuranose-phosphate								
$2-H_2PO_3$	10.48	0.58	0.36	2.04	1.42	0.58	0.62	0.95
$4-H_2PO_3$	10.80	0.75	0.03	2.31	1.69	0.66	0.80	1.08
MUE (vs DFT)		**0.67**	**0.29**	**2.12**	**1.50**	**0.65**	**0.69**	**1.02**
MSE (vs DFT)		**0.67**	**0.29**	**2.12**	**1.50**	**0.65**	**0.69**	**1.02**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5]. [b] DFT ionization obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $IP^{calc} - IP^{DFT}$.

Table A12. Absolute ionization potentials of β-D-carbohydrate conformers used in parameterization (eV)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose								
$^{1}C_{4}$	9.87	0.73	0.94	2.22	1.66	0.81	0.65	1.07
$^{4}C_{1}$	10.18	0.73	0.89	2.21	1.45	0.78	0.68	1.04
MUE (vs DFT)		**0.73**	**0.92**	**2.22**	**1.56**	**0.80**	**0.67**	**1.05**
MSE (vs DFT)		**0.73**	**0.92**	**2.22**	**1.56**	**0.80**	**0.67**	**1.05**
^{14}B	9.93	0.87	1.05	2.46	1.67	0.91	0.83	1.18
^{25}B	10.07	0.76	0.85	2.22	1.53	0.76	0.71	1.04
$^{2}S_{6}$	10.09	0.91	0.98	2.39	1.66	0.89	0.89	1.16
$^{3}S_{1}$	10.00	0.72	0.89	2.21	1.54	0.78	0.66	1.12
^{5}E	9.85	0.87	1.07	2.70	1.93	0.93	0.90	1.23
$^{5}S_{1}$	9.56	0.83	0.99	2.53	1.63	0.88	0.86	1.23
B_{14}	9.48	0.90	1.04	2.41	1.55	0.93	0.83	1.23
$^{1}S_{3}$	9.93	0.86	1.03	2.49	1.72	0.91	0.89	1.22
$^{2}H_{3}$	9.94	0.68	0.81	2.13	1.29	0.70	0.59	1.03
$^{6}H_{1}$	9.71	0.77	0.94	2.36	1.50	0.82	0.72	1.13
B_{36}	9.55	0.92	1.07	2.43	1.77	0.95	0.93	1.22
MUE (vs DFT)		**0.83**	**0.97**	**2.39**	**1.62**	**0.86**	**0.80**	**1.16**
MSE (vs DFT)		**0.83**	**0.97**	**2.39**	**1.62**	**0.86**	**0.80**	**1.16**
β-D-ribofuranose[a]								
1	10.21	0.69	0.69	1.94	1.24	0.65	0.49	0.87
2	9.95	0.77	0.83	2.19	1.40	0.76	0.69	1.01
3	10.18	0.59	0.64	1.93	1.14	0.55	0.39	0.80
4	9.87	0.78	0.89	2.36	1.56	0.79	0.78	1.09
5	9.96	0.71	0.81	2.15	1.49	0.69	0.59	0.88
6	10.17	0.71	0.80	2.16	1.50	0.68	0.61	0.90
6b	10.31	0.65	0.80	2.36	1.64	0.68	0.70	0.99
6ab	10.32	0.69	0.86	2.36	1.64	0.72	0.76	1.01
7	9.98	0.81	0.96	2.48	1.83	0.83	0.91	1.18
8	9.64	0.84	0.99	2.30	1.71	0.85	0.75	1.02
9	9.90	0.86	0.95	2.30	1.62	0.86	0.76	1.06
10	10.06	0.67	0.77	2.13	1.53	0.67	0.60	0.88
11	9.94	0.70	0.93	2.40	1.71	0.76	0.73	0.97
12	9.99	0.76	0.88	2.47	1.70	0.76	0.75	1.05
13	10.19	0.77	0.84	2.01	1.44	0.78	0.66	0.90
14	10.00	0.82	1.01	2.29	1.77	0.87	0.88	1.14
15	9.87	0.80	0.87	2.03	1.37	0.77	0.57	0.92
16	10.20	0.75	0.79	2.12	1.39	0.74	0.64	1.00
MUE (vs DFT)		**10.04**	**0.85**	**2.22**	**1.54**	**0.75**	**0.68**	**0.98**
MSE (vs DFT)		**10.04**	**0.85**	**2.22**	**1.54**	**0.75**	**0.68**	**0.98**

[a] Conformations obtained from ref 15 of Chapter 5. [b] DFT ionization obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $IP^{calc} - IP^{DFT}$.

Table A13. Absolute ionization potentials of α-D-carbohydrate conformers used in parameterization (eV)

	Reference	Error						
Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-D-glucopyranose								
$^1C_4$	9.77	0.64	0.90	2.44	1.69	0.74	0.67	1.07
$^4C_1$	10.13	0.76	0.78	1.98	1.23	0.69	0.52	0.84
MUE (vs DFT)		**0.70**	**0.84**	**2.21**	**1.46**	**0.72**	**0.60**	**0.95**
MSE (vs DFT)		**0.70**	**0.84**	**2.21**	**1.46**	**0.72**	**0.60**	**0.95**
^{14}B	9.99	0.76	0.91	2.43	1.58	0.79	0.75	1.10
^{25}B	9.73	0.84	0.99	2.50	1.77	0.88	0.86	1.17
$^2S_6$	10.13	0.74	0.91	2.63	1.76	0.80	0.87	1.16
$^3S_1$	10.16	0.74	0.88	2.22	1.58	0.79	0.67	1.12
$^5S_1$	9.99	0.71	0.92	2.40	1.73	0.79	0.75	1.12
B_{14}	9.91	0.81	0.99	2.43	1.70	0.87	0.82	1.20
$^1S_3$	9.80	0.88	0.99	2.53	1.63	0.90	0.90	1.22
$^2H_3$	9.99	0.78	0.83	2.01	1.28	0.75	0.53	1.01
$^6H_1$	9.57	0.83	0.96	2.22	1.47	0.85	0.67	1.10
B_{36}	9.71	0.83	0.96	2.46	1.75	0.85	0.86	1.13
MUE (vs DFT)		**0.79**	**0.93**	**2.38**	**1.63**	**0.83**	**0.77**	**1.13**
MSE (vs DFT)		**0.79**	**0.93**	**2.38**	**1.63**	**0.83**	**0.77**	**1.13**
α-D-ribofuranose[a]								
1	9.95	0.62	0.63	1.96	1.16	0.57	0.44	0.83
2	10.29	0.64	0.80	2.15	1.46	0.67	0.64	0.91
3	10.16	0.63	0.69	2.05	1.24	0.60	0.50	0.90
4	10.33	0.68	0.85	2.34	1.77	0.72	0.76	1.00
5	10.02	0.77	0.93	2.29	1.64	0.79	0.75	0.99
6	9.91	0.80	0.94	2.43	1.58	0.82	0.78	1.11
6b	9.90	0.80	0.94	2.44	1.59	0.82	0.78	1.11
6ab	9.95	0.82	0.95	2.47	1.60	0.84	0.81	1.14
7	9.88	0.80	0.89	2.48	1.54	0.80	0.84	1.14
8	9.76	0.64	0.81	2.08	1.49	0.66	0.53	0.79
9	10.13	0.75	0.86	2.16	1.54	0.75	0.66	0.94
10	9.66	0.87	0.94	2.31	1.44	0.87	0.75	1.11
11	9.83	0.84	0.90	2.24	1.45	0.82	0.72	1.09
12	10.23	0.79	0.90	2.37	1.67	0.80	0.79	1.06
13	9.96	0.87	0.88	2.31	1.47	0.84	0.83	1.09
14	9.87	0.85	0.95	2.48	1.60	0.86	0.91	1.17
15	9.86	0.86	0.91	2.17	1.47	0.83	0.69	1.01
16	9.92	0.87	0.95	2.34	1.51	0.87	0.78	1.18
MUE (vs DFT)		**0.77**	**0.87**	**2.28**	**1.51**	**0.77**	**0.72**	**1.03**
MSE (vs DFT)		**0.77**	**0.87**	**2.28**	**1.51**	**0.77**	**0.72**	**1.03**

[a] Initial conformations obtained from ref 15 of Chapter 5. [a] DFT ionization obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $IP^{calc} - IP^{DFT}$.

Table A14. Experimental and Theoretical interaction energies for molecules used in parameterization (kcal/mol)

	Reference		Error		
Molecule	Exp	DFT[b]	AM1	PM3	RM1
H_2O:H_2O	5.00[a]	-5.18	2.85	2.11	4.35
H_2O:CH_3OH		-5.17	3.18	2.45	4.40
H_2O:PO_3^-		-15.90	7.72	9.00	12.25
H_2O:$H_2PO_4^-$		-18.20	10.27	12.34	14.17
H_2O:HPO_4^{2-}		-33.27	15.95	10.51	15.71
MUE (vs DFT)			**7.99**	**7.28**	**10.18**
MSE (vs DFT)			**7.99**	**7.28**	**10.18**

[a] Experimental value obtained from Feyereisen et al.[ref 16 of Chapter 5] [b] The DFT interaction energies where computed with M06-2X/6-31+G(df). All errors are computed as $\Delta H_{\text{int}}^{calc} - \Delta H_{\text{int}}^{ref}$.

Table A15. Experimental and Theoretical Heats of formation for molecules used in parameterization (kcal/mol)

	Reference		Error						
Molecule	Exp	DFT[e]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
H_3O^+	138.0[a]	145.2	7.3	21.9	32.4	-99.1	2.9	6.2	3.5
H_2O	-57.8[a]	-55.3	-1.4	4.5	-44.9	-100.4	0	1.7	2.2
HO^-	-33.2[a]	-26.5	19.2	15.9	-118.5	-93.3	23.8	6.4	15.8
CH_3OH	-48.1[a]	-47.1	-7.6	-3.3	19.5	-181.7	-1.5	-2.9	-1.6
CH_3O^-	-36.0[a]	-29.6	-1.7	-1.4	-78.8	-185.5	7.1	2.8	-8.1
C_2H_5OH	-56.2[a]	-54.2	-4.6	0	117.5	-259.7	1.6	6.2	8.5
$C_2H_5O^-$	-47.5[a]	-40.0	2.9	3.3	25.8	-260.1	10.8	12.3	5.3
C_3H_7OH	-61.2[c]	-58.4	-5.6	-0.1	211.1	-340.6	1.9	12.0	15.6
C_4H_9OH	-66.2[c]	-62.4	-6.7	-0.2	304.9	-421.2	2.3	18.0	22.8
C_6H_5OH	-23.0[a]	-22.9	2.3	2.2	128.2	-758.8	0.7	8.2	25.0
$C_6H_5O^-$	-40.5[a]	-39.5	0.2	-3.1	44.7	-771.8	1.4	6.3	23.2
CH_3CO_2H	-103.3[a]	-103.7	4.5	3.6	-156.7	-443.9	3.6	12.9	-6.0
$CH_3CO_2^-$	-122.5[a]	-121.8	9.4	4.4	-242.3	-451.0	9.0	7.1	-9.9
CH_3OCH_3	-43.99[c]	-44.1	-6.6	-3.6	91.1	-256.8	-0.2	0.1	-2.8
MUE (vs exp)		**3.0**	**5.7**	**4.8**	**115.5**	**330.3**	**4.8**	**7.4**	**10.7**
MSE (vs exp)		**2.9**	**0.8**	**3.2**	**23.9**	**-330.3**	**4.5**	**7.0**	**6.7**
$P(CH_3)_3$	-24.2[a]	-24.1	14.3	-4.7	406.8	-158.7	-13.5	16.0	-11.1
$(CH_3)_3PO$	-103.8[a]	-95.7	27.5	22.2	239.2	-281.9	8.3	4.4	1.9
PO_3^-	-225.4[a]	-221.3	22.3	28.7	-605.6	-404.7	29.7	19.7	10.9
H_3PO_4	-272.8[d]	-272.2	-10.2	15.7	-606.6	-537.9	8.9	34.8	14.6
MUE (vs exp)		**3.2**	**18.6**	**17.8**	**464.6**	**345.8**	**15.1**	**18.7**	**9.6**
MSE (vs exp)		**3.2**	**13.5**	**15.5**	**-141.6**	**-345.8**	**8.4**	**18.7**	**4.1**
Aspartic acid ($C_4NH_7O_4$)	-185.9[b]	-192.4	0.8	6.9	-338.2	-924.7	1.3	20.1	-13.7
Asparagine ($C_4N_2H_8O_3$)	-137.8[b]	-145.6	3.4	11.1	-103.1	-737.6	-4.1	22.7	-1.8
Glutamic acid ($C_5NH_9O_4$)	-189.7[b]	-196.5	-1.5	4.0	-247.2	-1007.0	-0.7	23.5	-8.9
Glutamine ($C_5N_2H_{10}O_3$)	-142.4[b]	-148.3	3.0	9.7	-10.8	-819.7	-4.2	28.5	4.3
Histidine ($C_6N_3H_9O_2$)	-63.5[b]	-67.1	28.6	11.7	117.9	-836.0	1.5	45.3	-0.5
Argenine ($C_6N_4H_{14}O_2$)	-84.3[b]	-92.2	14.9	11.1	374.7	-715.4	7.4	52.8	23.5
Phenylalanine ($C_9NH_{11}O_2$)	-69.3[b]	-74.9	-1.1	2.3	194.9	-1157.2	-1.7	22.0	30.0
Tyrosine ($C_9NH_{11}O_3$)	-111.6[b]	-118.9	-2.4	0.1	32.7	-1326.0	-4.9	18.5	19.3
Tryptophan ($C_{11}N_2H_{12}O_2$)	-51.6[b]	-57.8	20.3	8.6	257.1	-1397.4	-2.5	42.3	33.2
MUE (vs exp)		**6.4**	**8.4**	**7.3**	**186.3**	**991.2**	**3.1**	**30.6**	**15.0**
MSE (vs exp)		**-6.4**	**7.3**	**7.3**	**30.9**	**-991.2**	**-0.9**	**30.6**	**9.5**

[a, b] Experimental values obtained from Nam et al.[ref 8 of Chapter 5] [c] Experimental values obtained from NIST chemistry webbook.[ref 14 of Chapter 5] [d] Theoretical values derived by Alexeev et al.[ref 17 of Chapter 5] [e] The DFT heats of formation where computed with M06-2X/6-311++G(3df,2p) level of theory and basis set. All errors are computed as $\Delta H_{\text{int}}^{calc} - \Delta H_{\text{int}}^{\text{exp}}$.

Table A16. Relative heats of formation for different protonated forms of carbohydrate-phosphate conformers used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose-phosphate								
$^1C_4$–H_2PO_4	4.20	0	7.83	22.51	15.40	10.97	0	0.56
$^4C_1$–H_2PO_4	0	0.44	0	0	0	0	9.33	0
$^1C_4$–HPO_4^-	0	0	0	0	0	0	0	0
$^4C_1$–HPO_4^-	0.02	9.68	12.32	3.87	9.46	3.61	15.44	4.37
$^1C_4$–PO_4^{2-}	0	0	0	0	0	0	0	0
$^4C_1$–PO_4^{2-}	12.19	9.28	14.57	12.67	13.38	25.92	23.93	30.21
MUE (vs DFT)		**2.87**	**3.05**	**3.77**	**3.64**	**4.02**	**6.78**	**4.33**
MSE (vs DFT)		**0.50**	**3.05**	**3.77**	**3.64**	**4.02**	**5.38**	**3.12**
β-D-ribofuranose-phosphate[a]								
6–H_2PO_4	0	1.78	2.54	2.57	2.24	2.98	2.10	0
12–H_2PO_4	1.77	0	0	0	0	0	0	0.49
6–HPO_4^-	0	4.35	5.75	3.12	3.73	3.96	6.35	1.66
12–HPO_4^-	1.64	0	0	0	0	0	0	0
6–PO_4^{2-}	1.54	7.80	14.34	16.75	17.01	4.19	7.18	5.61
12–PO_4^{2-}	0	0	0	0	0	0	0	0
MUE (vs DFT)		**2.63**	**4.08**	**4.05**	**4.14**	**2.17**	**2.92**	**1.44**
MSE (vs DFT)		**1.50**	**2.95**	**2.92**	**3.01**	**1.03**	**1.78**	**0.47**
α-D-glucopyranose-phosphate								
$^1C_4$–H_2PO_4	8.11	7.46	0	6.66	1.72	3.14	4.68	9.66
$^4C_1$–H_2PO_4	0	0	1.16	0	0	0	0	0
$^1C_4$–HPO_4^-	3.73	9.32	0	0	0	0	3.86	0
$^4C_1$–HPO_4^-	0	0	5.69	4.25	6.97	5.75	0	3.18
$^1C_4$–PO_4^{2-}	2.11	0	0	0	0	0	0	0
$^4C_1$–PO_4^{2-}	0	8.24	21.04	21.65	17.90	12.10	3.22	5,89
MUE (vs DFT)		**2.77**	**6.97**	**5.53**	**6.18**	**4.78**	**1.48**	**2.74**
MSE (vs DFT)		**1.85**	**2.32**	**3.10**	**2.11**	**1.17**	**-0.37**	**0.80**
α-D-ribofuranose-phosphate								
2–H_2PO_4	2.92	0	0	0	0	0	5.21	4.66
4–H_2PO_4	0	2.01	3.71	8.78	4.63	2.84	0	0
2–HPO_4^-	3.14	0	0	0	0	1.01	5.81	4.22
4–HPO_4^-	0	1.88	1.83	2.66	0.47	0	0	0
2–PO_4^{2-}	8.18	7.69	19.45	24.25	20.66	8.40	6.72	3.93
4–PO_4^{2-}	0	0	0	0	0	0	0	0
MUE (vs DFT)		**1.74**	**3.81**	**5.60**	**3.94**	**1.35**	**1.07**	**1.18**
MSE (vs DFT)		**-0.44**	**1.79**	**3.58**	**1.92**	**-0.33**	**0.58**	**-0.24**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table A17. Relative heats of formation for β-D-carbohydrate conformers used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose								
$^1C_4$	4.8	0	0.2	10.2	5.7	1.6	0	5.2
$^4C_1$	0	1.7	0	0	0	0	7.2	0
MUE (vs DFT)		**3.3**	**2.3**	**2.7**	**0.5**	**1.6**	**6.0**	**0.2**
MSE (vs DFT)		**-1.6**	**-2.3**	**2.7**	**0.5**	**-1.6**	**1.2**	**0.2**
^{14}B	7.1	6.3	7.6	13.5	11.7	5.8	5.8	8.8
^{25}B	15.7	17.6	15.6	21.4	19.7	16.9	17.7	18.7
$^2S_6$	0.2	3.1	3.2	4.7	2.8	3.1	4.1	4.1
$^3S_1$	3.9	5.3	6.2	13.6	10.1	6.7	7.3	8.7
^5E	4.2	4.2	3.8	9.6	8.5	8.0	1.4	9.2
$^5S_1$	3.4	7.8	8.2	7.7	5.7	6.3	6.9	2.6
B_{14}	10.5	13.6	14.7	18.8	14.0	14.3	14.7	14.2
$^1S_3$	1.7	0	0	0	0	0	0	0
$^2H_3$	5.6	12.5	12.0	9.5	9.2	12.1	17.2	10.1
$^6H_1$	0	5.5	5.3	2.6	1.4	5.0	10.0	2.6
B_{36}	13.4	17.0	16.7	15.1	15.7	14.9	17.6	16.1
MUE (vs DFT)		**2.9**	**2.9**	**4.9**	**3.3**	**3.0**	**4.4**	**3.1**
MSE (vs DFT)		**2.5**	**2.5**	**4.6**	**3.0**	**2.5**	**3.4**	**2.7**
β-D-ribofuranose[a]								
1	0.7	4.8	6.1	0.6	2.3	2.7	6.4	0.1
2	4.3	7.7	7.0	2.8	3.4	6.0	11.0	3.8
3	2.2	2.3	3.4	1.1	1.4	0.8	5.2	0.6
4	3.9	5.5	4.7	1.9	1.8	3.9	8.4	2.9
5	0.7	1.1	2.5	1.3	1.2	0	3.6	0
6	0	2.1	3.5	1.8	1.7	2.2	3.6	1.7
6b	0.8	1.4	3.7	5.8	4.3	2.0	4.4	3.4
6ab	1.6	1.6	3.8	5.4	3.8	1.5	4.4	2.5
7	1.3	0	0	0.5	0	0.9	0	1.2
8	4.0	1.3	2.5	3.4	3.0	1.7	3.5	2.1
9	2.2	1.6	3.0	1.7	1.9	2.1	2.4	1.5
10	1.4	4.5	6.1	5.2	4.3	4.7	5.5	4.0
11	0.6	2.4	3.5	2.4	1.6	1.7	3.0	1.0
12	0.6	0.6	2.3	1.0	1.0	2.3	0.4	1.7
13	4.5	2.6	1.9	0	0.8	1.2	3.7	1.0
14	4.9	1.6	0.6	0.9	0.9	0.2	2.7	1.5
15	4.8	6.0	5.4	3.0	4.8	2.9	9.4	2.3
16	1.3	3.6	4.5	1.0	2.2	1.5	5.5	0.4
MUE (vs DFT)		**1.7**	**2.5**	**1.9**	**1.6**	**1.6**	**3.0**	**1.5**
MSE (vs DFT)		**0.6**	**1.4**	**0**	**0**	**-0.1**	**2.4**	**-0.5**

[a] Conformations obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table A18. Relative Heats of formation for α-D-carbohydrate conformers used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-D-glucopyranose								
$^1C_4$	2.9	0.7	0	5.1	3.4	1.3	0	3.0
$^4C_1$	0	0	0.8	0	0	0	0.9	0
MUE (vs DFT)		**1.1**	**1.9**	**1.1**	**0.3**	**0.8**	**1.9**	**0.04**
MSE (vs DFT)		**-1.1**	**-1.1**	**1.1**	**0.3**	**-0.8**	**-1.0**	**0.04**
^{14}B	5.2	2.3	2.7	7.6	7.9	0.5	6.3	4.7
^{25}B	17.8	17.0	14.7	20.2	19.7	14.4	19.6	18.1
$^2S_6$	1.9	0	0	0.6	2.1	0.4	0.6	3.0
$^3S_1$	1.9	4.1	3.6	11.3	7.1	3.5	4.0	7.7
$^5S_1$	0	2.5	1.6	1.8	0	0	0	0.0
B_{14}	3.5	8.3	4.6	8.6	2.7	3.9	6.5	6.2
$^1S_3$	2.9	3.1	2.2	0.5	0.1	2.8	4.4	3.2
$^2H_3$	6.8	12.1	10.6	5.3	8.4	7.8	17.1	7.3
$^6H_1$	1.4	7.1	7.0	0	1.6	3.3	11.3	1.3
B_{36}	10.0	12.2	9.9	6.2	8.8	7.2	14.3	8.9
MUE (vs DFT)		**2.9**	**2.2**	**3.2**	**1.7**	**1.7**	**3.5**	**1.2**
MSE (vs DFT)		**1.7**	**0.6**	**1.1**	**0.7**	**-0.8**	**3.3**	**0.9**
α-D-ribofuranose[a]								
1	5.5	10.5	9.9	7.0	6.6	7.8	16.1	6.4
2	0.2	2.6	3.6	4.0	2.8	2.3	5.9	3.8
3	1.5	2.3	3.6	1.5	0.5	1.7	5.4	1.7
4	0	0.5	1.5	3.5	1.6	0.4	3.5	3.2
5	0.3	1.0	2.3	0.3	0	0	2.3	0
6	0.5	2.4	2.6	1.6	0.1	1.2	5.0	2.2
6b	0.6	2.9	3.1	2.2	0.8	1.4	5.6	2.5
6ab	1.2	3.2	3.3	2.0	0.4	1.2	5.5	1.7
7	5.2	7.0	3.2	7.5	4.2	5.7	7.6	6.9
8	6.2	3.0	3.3	2.7	3.9	0.4	5.4	1.0
9	0.5	1.6	1.9	0	1.6	0.9	1.7	0.9
10	2.7	6.5	6.3	4.3	2.9	4.6	9.2	4.6
11	2.1	4.7	5.3	2.0	2.1	2.9	6.0	1.8
12	0.9	0	0	1.0	1.8	0	0	1.7
13	6.0	8.2	4.1	5.9	3.1	6.4	10.2	7.1
14	6.4	7.1	2.5	6.6	2.9	5.2	8.9	7.5
15	4.9	6.1	5.7	3.3	4.0	4.4	10.1	3.9
16	1.0	3.8	4.5	2.6	1.0	2.1	7.3	2.3
MUE (vs DFT)		**2.0**	**2.5**	**1.3**	**1.2**	**1.1**	**4.1**	**1.5**
MSE (vs DFT)		**1.5**	**1.2**	**0.7**	**-0.3**	**0.2**	**3.9**	**0.7**

[a] Conformations obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table A19. Relative Heats of formation for dihedral scans used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1	PM3	PM3CARB-1	PM3^{MS}	RM1	AM1/d-PhoT	AM1/d-CB1
1, 2-Ethanediol[a]								
Dihedral C-C-O-H								
1	0.001	0.000	0.000	0.016	0.014	0.003	0.002	0.007
2	2.780	0.001	0.005	0.012	0.017	0.005	0.001	0.007
3	3.234	3.854	2.435	1.777	2.028	3.307	5.362	2.425
4	2.971	3.852	2.434	1.790	2.035	3.306	5.362	2.427
5	0.224	0.000	0.002	0.000	0.004	0.002	0.000	0.003
6	1.546	0.001	0.000	0.000	0.000	0.000	0.005	0.000
7	0.000	0.002	0.005	0.016	0.015	0.007	0.004	0.009
MUE (vs DFT)		**0.865**	**0.841**	**1.029**	**0.957**	**0.709**	**1.296**	**0.844**
MSE (vs DFT)		**-0.435**	**-0.839**	**-1.021**	**-0.949**	**-0.589**	**-0.003**	**-0.840**
Dihedral O-C-C-O								
1	0.000	0.000	0.000	0.000	0.063	0.000	0.000	0.097
2	4.312	0.464	0.014	0.038	0.144	0.156	0.011	0.000
3	0.479	0.471	0.024	0.113	0.153	0.163	0.042	0.003
4	5.319	0.472	0.019	0.084	0.141	0.159	0.044	0.000
5	2.645	1.385	0.324	0.818	0.340	0.774	2.301	1.024
6	6.018	1.389	0.331	0.780	0.321	0.769	2.307	1.014
7	2.273	1.177	0.143	0.478	0.000	0.590	1.076	0.622
MUE (vs DFT)		**2.241**	**2.885**	**2.677**	**2.859**	**2.634**	**2.181**	**2.640**
MSE (vs DFT)		**-2.241**	**-2.885**	**-2.677**	**-2.841**	**-2.634**	**-2.181**	**-2.612**
Methoxymethanol[a]								
Dihedral O-C-O-H								
1	0.001	0.004	0.004	0.007	0.005	0.004	0.003	0.006
2	1.977	0.000	0.001	0.000	0.000	0.000	0.000	0.000
3	2.998	6.557	4.030	5.954	4.378	6.309	8.619	6.382
4	2.909	0.008	0.006	0.011	0.007	0.006	0.006	0.008
5	0.000	0.000	0.000	0.009	0.003	0.000	0.000	0.000
MUE (vs DFT)		**1.688**	**1.183**	**1.570**	**1.254**	**1.639**	**2.101**	**1.654**
MSE (vs DFT)		**-0.263**	**-0.769**	**-0.381**	**-0.698**	**-0.313**	**0.149**	**-0.298**
Dihedral C-O-C-O								
1	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.001
2	1.448	0.006	0.003	0.003	0.000	0.003	0.009	0.000
3	2.974	0.005	0.003	0.009	0.007	0.004	0.005	0.003
4	3.060	3.039	1.947	1.860	2.306	3.704	5.803	3.031
5	2.233	3.062	1.956	1.854	2.289	3.716	5.837	3.042
6	2.885	3.035	1.946	1.869	2.321	3.705	5.791	3.039
7	1.782	2.948	2.277	3.972	2.324	3.116	4.405	3.431
MUE (vs DFT)		**0.940**	**1.035**	**1.314**	**0.905**	**1.242**	**2.326**	**1.009**
MSE (vs DFT)		**-0.327**	**-0.893**	**-0.688**	**-0.733**	**-0.019**	**1.067**	**-0.263**

[a] Conformations used were determined as maxima and minima along the dihedral scanned potential energy surfaces. [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Appendix B

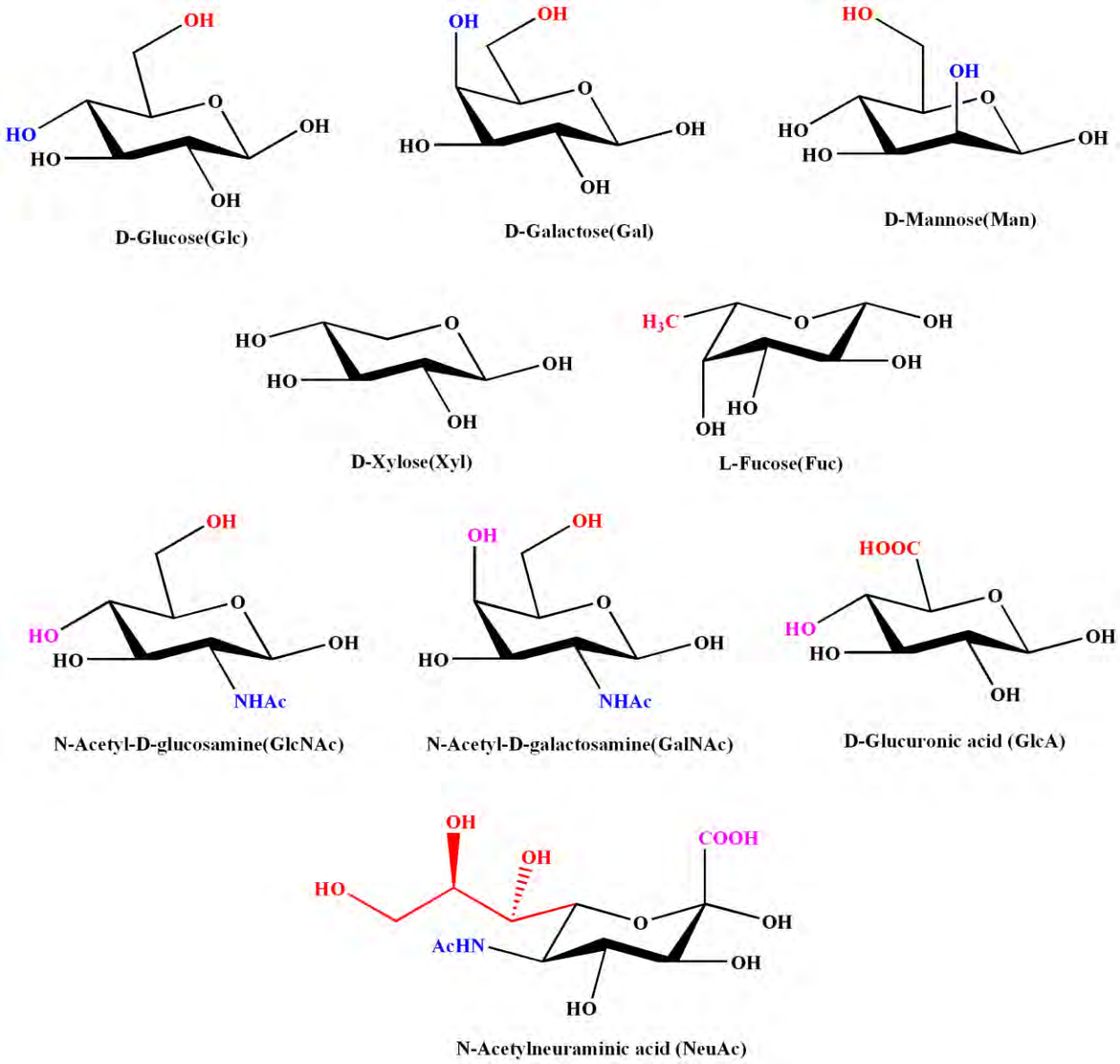


Figure B1: Molecular structures of nine monosaccharaides important in mammalian cells.

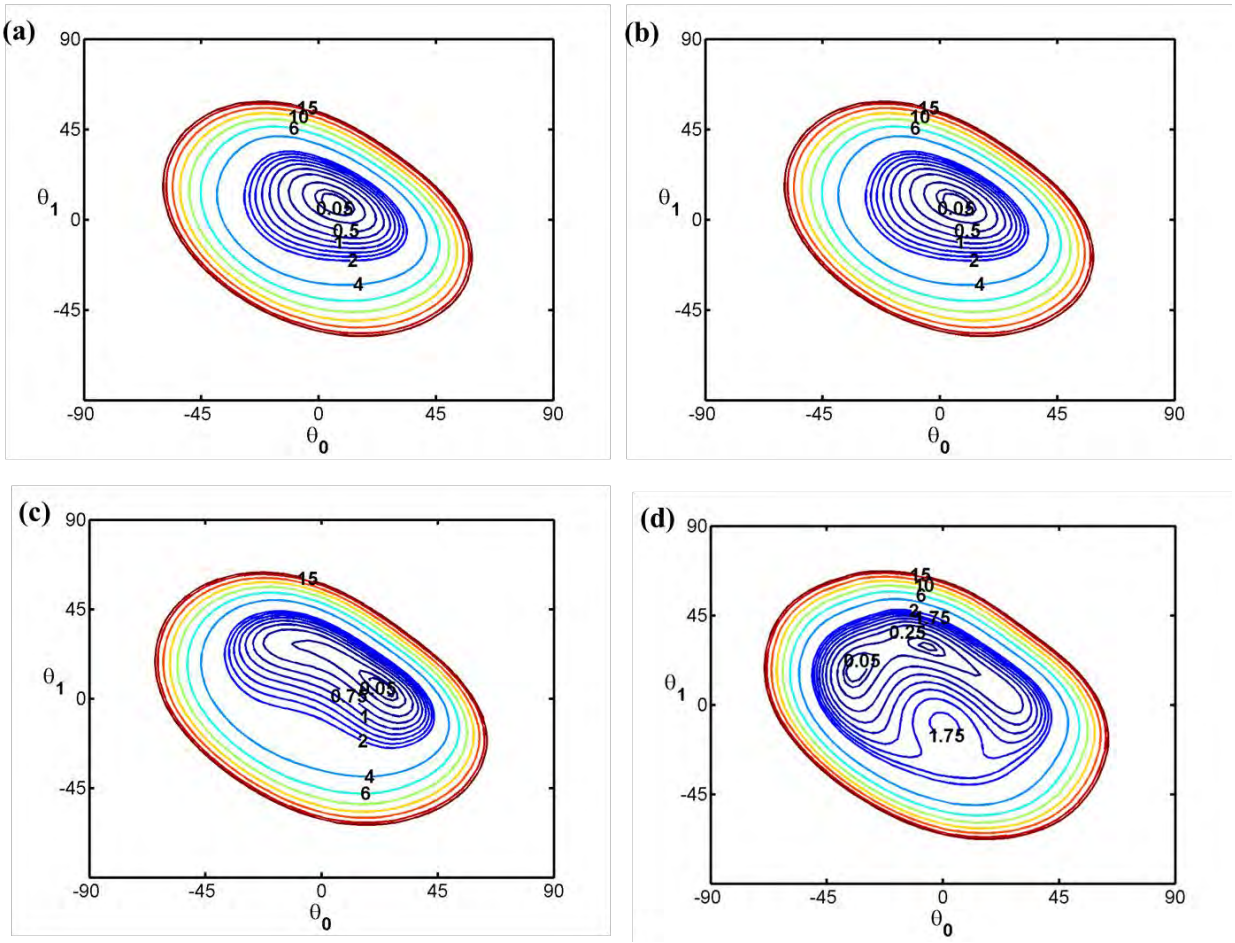


Figure B2: Ribofuranose free energy of puckering shown as two-dimensional contour plots for (a) AM1, (b) PM3, (c) RM1 and (d) SCC-DFTB. Energy is mapped to color from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.05 kcal/mol to 0.1 kcal/mol, then from 0.1 kcal/mol to 2 kcal/mol in steps of 0.25 kcal/mol and every 2 kcal/mol thereafter.

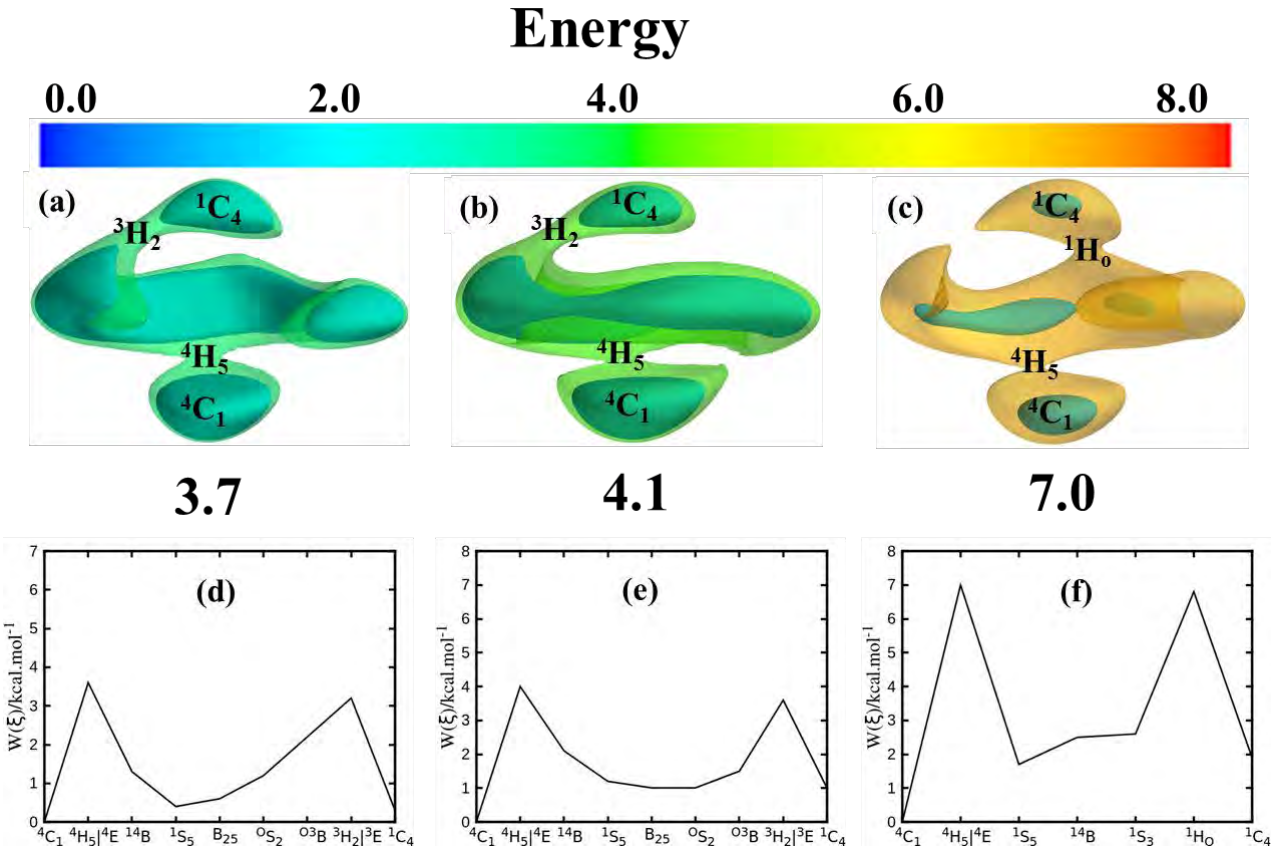


Figure B3: Free energy of puckering color mapped to three-dimensional volumes of glucopyranose for (a) AM1, (b) PM3 and (c) RM1. Volumes are mapped to color from 0 kcal/mol (blue) to 8 kcal/mol (red). The inner isosurface is at 3 kcal/mol and the outer isosurface indicates the minimum free energy to connect the $^1C_4$ and $^4C_1$ conformers which occurs at 3.7, 4.1 and 7.0 kcal/mol, respectively. The one-dimensional minimum free energy paths have been extracted from the free energy volumes and are shown for (d) AM1, (e) PM3 and (f) RM1.

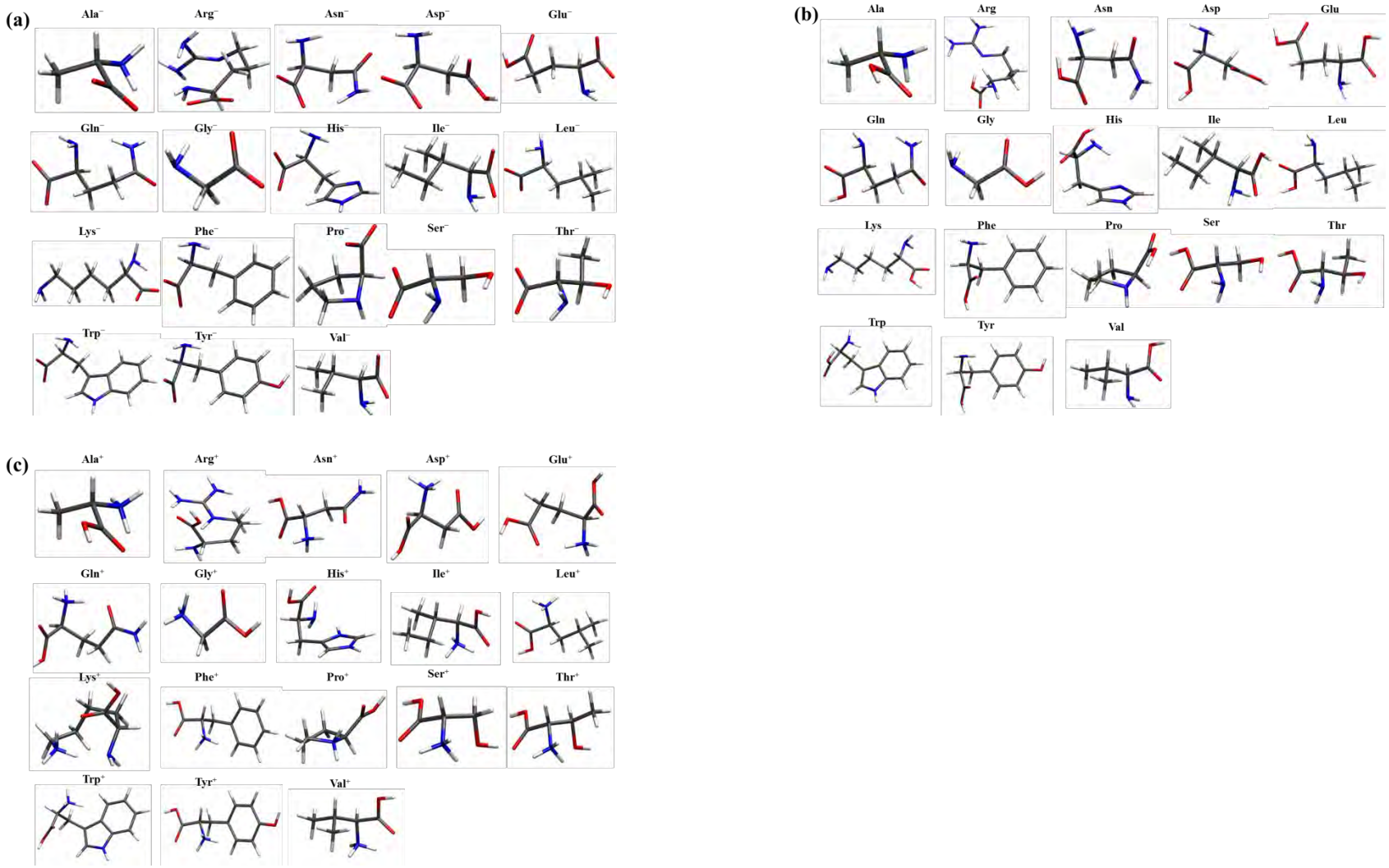


Figure B4: Molecular structures of various amino acids provided with their three letter abbreviations in (a) anionic form, (b) neutral form, and (c) cationic form. Abbreviations correspond to: Ala = Alanine, Arg = Arginine, Asn = Asparagine, Asp = Aspartic acid, Glu = Glutamic acid, Gln = Glutamine, Gly = Glycine, His = Histidine, Ile = Isoleucine, Leu = Leucine, Lys = Lysine, Phe = Phenylalanine, Pro = Proline, Ser = Serine, Thr = Threonine, Trp = Tryptophan, Tyr = Tyrosine, and Val = Valine.

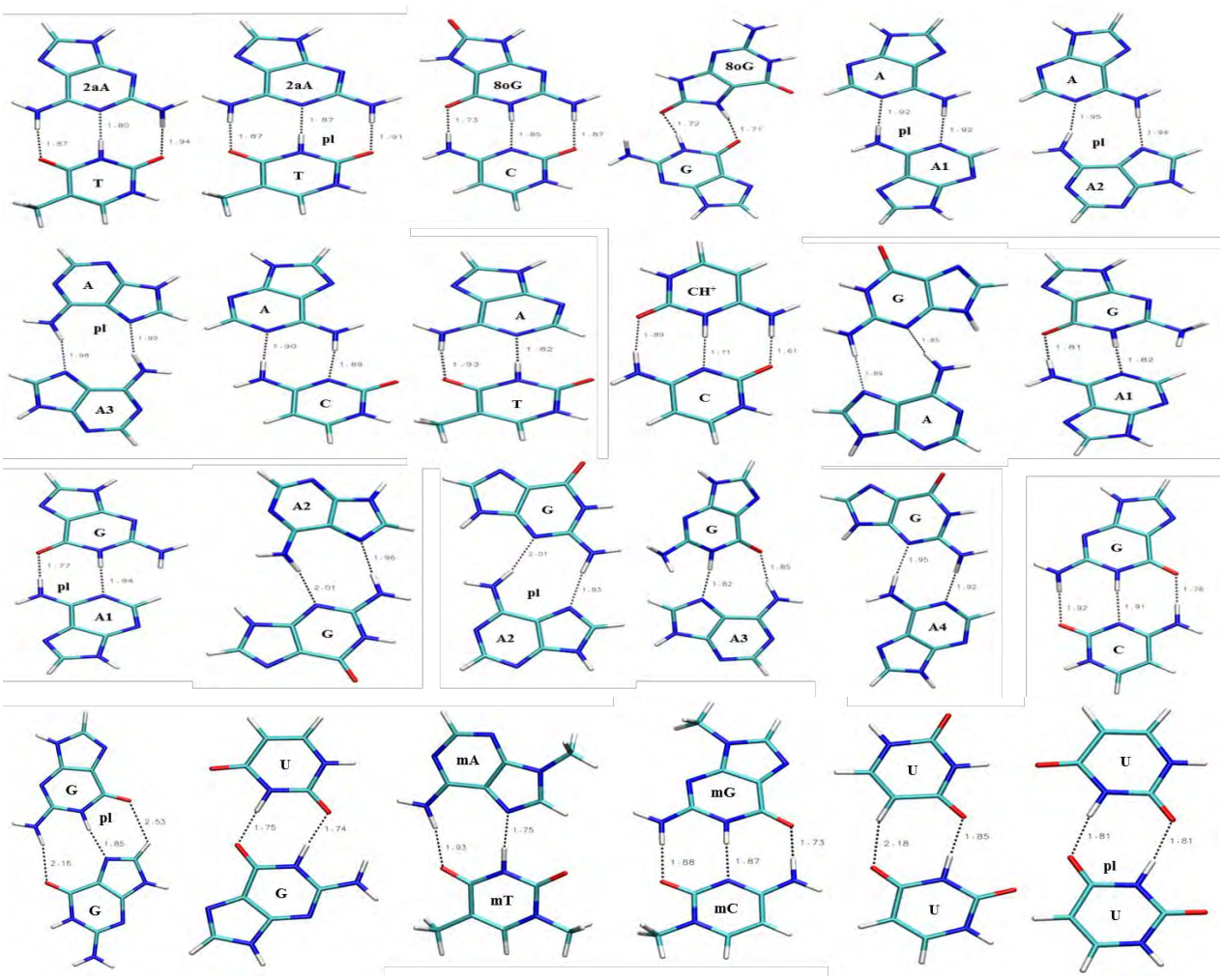


Figure B5: Hydrogen bonded DNA base pairs, along with hydrogen bond distances, contained in the JSCH-2005 dataset and utilized in the current work. Abbreviations used: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, a = amino, m = methyl, o = oxy, and pl = planar.

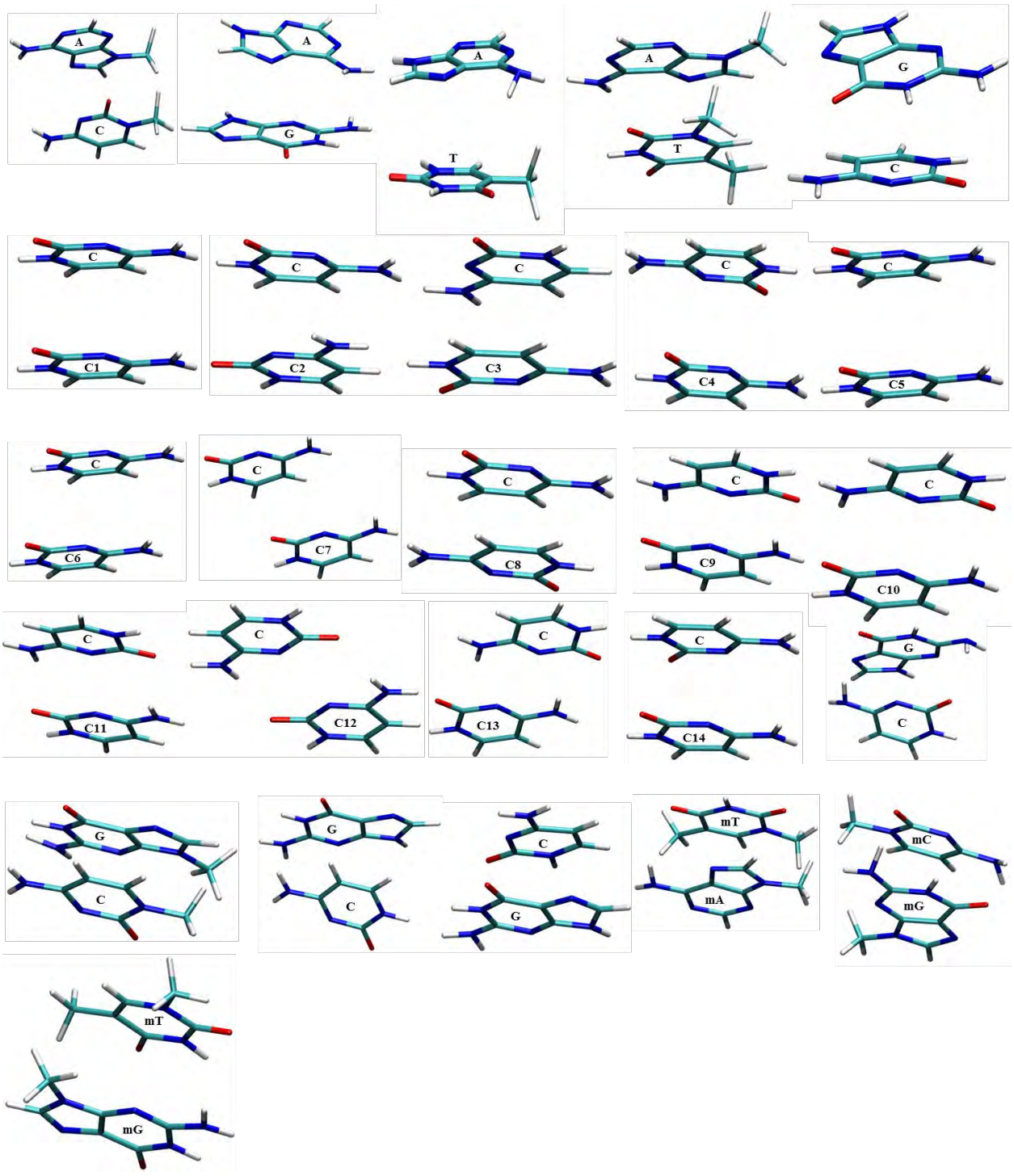


Figure B6: Stacked DNA base pairs contained in the JSCH-2005 dataset and utilized in the current work. Abbreviations used: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, a = amino, m = methyl, and o = oxy.

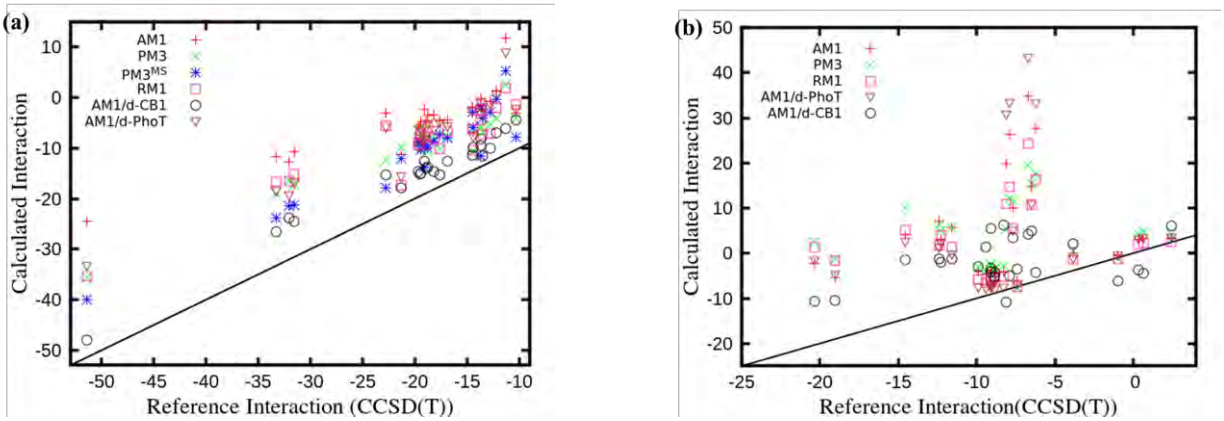


Figure B7: Gas phase interaction energies of various, (a) hydrogen bonded, and (b) stacked base pairs. The reference interaction energies are from CCSD(T) simulations. Methods used in the calculated interaction energies are indicated by various labels, and units for the energies are all in kcal/mol. All calculated energies were acquired via single point calculations.

Table B1: Relative heats of formation for nine monosaccharaides obtained from single point calculations (kcal/mol)

Molecule	DFT[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-L-Fucose	40.58	45.66	36.97	202.08	196.97	38.51	50.16	50.13
α-N-Acetylneuraminic acid	-144.25	-111.68	-119.07	-288.01	-1019.71	-129.79	-97.44	-145.80
β-D-Galactose	2.28	4.61	3.67	3.02	3.49	4.16	4.52	3.52
β-N-Acetyl-D-galactosamine	-0.91	20.72	7.76	145.98	-150.83	5.80	33.07	0.92
β-N-Acetyl-D-glucosamine	-7.18	17.95	6.44	140.56	-155.61	1.22	27.85	-4.75
β-D-Glucose	0	0	0	0	0	0	0	0
β-D-Glucuronic acid	-37.43	-27.80	-37.85	-312.35	-223.00	-37.60	-23.81	-51.57
β-D-Mannose	4.12	0.17	-0.92	8.18	5.41	1.00	2.30	6.09
β-D-Xylose	46.63	49.66	42.72	105.64	278.45	45.56	49.19	46.66
MSE (vs DFT)		**10.61**	**3.99**	**11.25**	**-107.63**	**2.78**	**15.78**	**0.15**
MUE (vs DFT)		**11.48**	**6.87**	**104.29**	**194.45**	**4.21**	**16.18**	**3.64**

[a] DFT energies were computed with M06-2X/6-311++G(3df,2p). All SE simulations were conducted as single point calculations on the DFT optimized structures. All errors are computed as $\Delta H_f^{calc} - \Delta H_f^{DFT}$.

Table B2: Relative heats of formation for nine monosaccharaides obtained from geometry optimization calculations (kcal/mol)

Molecule	DFT[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-L-Fucose	40.58	45.31	37.72	187.44	198.97	37.88	51.65	49.53
α-N-Acetylneuraminic acid	-144.25	-117.92	-122.95	-338.72	-1026.48	-134.58	-113.20	-153.11
β-D-Galactose	2.28	4.23	4.60	-15.31	1.97	2.37	5.71	-1.07
β-N-Acetyl-D-galactosamine	-0.91	17.46	5.57	110.43	-160.53	-0.23	28.84	-1.88
β-N-Acetyl-D-glucosamine	-7.18	10.85	4.31	116.63	-153.75	-1.25	18.39	-6.91
β-D-Glucose	0	0	0	0	0	0	0	0
β-D-Glucuronic acid	-37.43	-29.25	-37.91	-321.52	-226.24	-38.41	-27.64	-56.20
β-D-Mannose	4.12	-0.15	-0.93	-9.30	3.07	-0.54	2.66	2.77
β-D-Xylose	46.63	49.32	43.75	98.12	276.44	45.11	51.24	44.00
MSE (vs DFT)		**8.45**	**3.37**	**-8.45**	**-110.04**	**0.72**	**12.65**	**-2.97**
MUE (vs DFT)		**9.39**	**5.87**	**104.79**	**196.31**	**2.91**	**12.97**	**5.02**

[a] DFT energies were computed with M06-2X/6-311++G(3df,2p). All SE simulations were conducted as geometry optimizations on the DFT optimized structures. All errors are computed as $\Delta H_f^{calc} - \Delta H_f^{DFT}$.

Table B3: Root mean square deviations for nine monosaccharaides

Molecule	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
α-L-Fucose	0.043926	0.056197	0.180661	0.095900	0.044165	0.092648	0.102910
α-N-Acetylneuraminic acid	0.069833	0.073830	0.222803	0.114554	0.051427	0.113805	0.089155
β-D-Galactose	0.058514	0.058791	0.136011	0.075196	0.065091	0.100784	0.078731
β-N-Acetyl-D-galactosamine	0.064786	0.070104	0.219161	0.118476	0.136251	0.103369	0.081573
β-N-Acetyl-D-glucosamine	0.149202	0.122576	0.443240	0.099587	0.141875	0.256777	0.152352
β-D-Glucose	0.091453	0.128578	0.362342	0.066088	0.083627	0.118820	0.070466
β-D-Glucuronic acid	0.041274	0.057494	0.120968	0.145979	0.043564	0.116661	0.090185
β-D-Mannose	0.039051	0.065362	0.110686	0.070903	0.079123	0.075595	0.047531
β-D-Xylose	0.035961	0.048890	0.136626	0.067519	0.043222	0.068400	0.108763

$RMSD = \sqrt{\sum_{c=1}^{N} \left(calc_c - ref_c \right)^2 / N}$, where c represents the corresponding coordinate (x, y, or z), calc is the SE coordinate and ref is the M06-2X/6-311++G(3df,2p) coordinate, and N represents the total number of coordinates per molecule.

Table B4: Experimental and theoretical gas phase proton affinities for various amino acids obtained from geometry optimization (kcal/mol)

	Reference	Error							
Amino acid	Exp[a]	G3MP2[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
Reaction 1: $A + H^+ = AH^+$									
Alanine	215.5	0	-10.5	-9.1	-76.0	-34.6	0.3	-4.2	-3.1
Arginine	251.2	-1.1	-5.1	-24.2	-78.0	-29.6	1.9	7.9	0.2
Asparagine	222.0	1.6	-4.7	-4.0	-67.2	-27.7	10.4	0.1	4.9
Aspartic acid	217.2	1.8	-10.1	-9.8	-84.3	-22.9	2.8	-8.8	-0.3
Glutamic acid	218.2	8.3	-7.7	-7.1	-68.9	-12.8	5.3	1.4	5.9
Glutamine	224.1	8.5	-8.7	-7.2	-60.2	-15.7	7.1	7.8	11.0
Glycine	211.9	0	-12.3	-11.0	-74.4	-29.6	-0.1	-4.0	-2.6
Histidine	236.1	-2.2	-7.3	-11.4	-82.5	-34.9	4.4	0.3	8.4
Isoleucine	219.3	0.2	-12.2	-13.2	-73.6	-27.8	-0.8	-4.4	-3.9
Leucine	218.6	-0.1	-7.7	-13.2	-81.6	-31.5	-0.3	-3.2	-1.9
Lysine	238.0	1.1	-13.9	-17.7	-78.5	-25.4	0.7	1.2	12.2
Phenylalanine	220.6	0.6	-10.8	-13.3	-94.1	-40.2	-2.8	-2.4	-6.1
Proline	220.0	5.1	-0.5	-12.2	-65.3	-25.9	2.1	3.4	-5.2
Serine	218.6	-0.5	-12.7	-15.0	-68.4	-28.7	-3.4	-4.1	-3.6
Threonine	220.5	-0.8	-9.5	-11.0	-62.3	-27.1	-1.2	-2.9	-9.4
Tryptophan	226.8	-2.1	-5.1	-8.2	-64.4	-37.4	-0.5	0.7	-5.9
Tyrosine	221.0	1	-10.8	-12.7	-75.0	-15.1	-0.1	-3.4	-3.3
Valine	217.6	0.9	-6.8	-10.8	-77.2	-27.5	-0.2	-3.2	-4.6
MSE (vs. Exp)		**1.2**	**-8.7**	**-11.7**	**-74.0**	**-27.5**	**1.4**	**-1.0**	**-0.4**
MUE (vs. Exp)		**2.0**	**8.7**	**11.7**	**74.0**	**27.5**	**2.5**	**3.5**	**5.1**
	CBS-QB3[c]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT		AM1/d-CB1
Reaction 2: $A^- + H^+ = AH$									
Alanine	332.7	16.1	8.2	-74.4	7.5	17.6	8.5		6.8
Arginine	339.2	7.8	4.1	-70.8	-1.5	14.9	2.2		3.7
Asparagine	322.0	14.1	5.2	-78.2	-1.6	12.8	0.0		-1.4
Aspartic acid	328.9	13.8	9.2	-63.1	2.9	13.7	9.9		1.2
Glutamic acid	330.5	15.8	10.4	-66.5	6.7	17.5	6.9		7.4
Glutamine	324.3	17.4	11.8	-63.7	2.5	16.6	6.1		0.1
Glycine	332.6	14.3	8.4	-60.7	1.8	13.9	6.1		3.6
Histidine	332.5	4.1	6.3	-70.5	-2.1	11.2	-3.8		11.9
Isoleucine	330.3	16.8	11.8	-87.2	-1.2	19.7	9.8		6.3
Leucine	331.2	16.3	9.4	-75.0	2.4	18.0	8.5		6.3
Lysine	330.0	16.2	9.1	-66.4	1.2	17.3	4.8		6.0
Phenylalanine	328.4	14.3	11.5	-85.6	2.4	18.2	5.7		10.5
Proline	338.4	7.5	7.8	-72.8	1.8	11.5	1.1		0.8
Serine	326.5	17.4	13.2	-77.9	4.6	16.0	7.6		3.0
Threonine	325.8	18.3	10.9	-59.8	9.1	19.1	9.0		9.6
Tryptophan	329.5	5.2	9.0	-116.6	4.1	13.0	1.0		6.1
Tyrosine	328.7	15.8	12.3	-25.9	6.2	18.3	11.9		7.5
Valine	330.5	17.4	11.7	-66.7	5.4	18.3	10.0		8.3
MSE (vs. Exp)		**13.8**	**9.5**	**-71.2**	**2.9**	**16.0**	**5.8**		**5.4**
MUE (vs. Exp)		**13.8**	**9.5**	**71.2**	**3.6**	**16.0**	**6.3**		**5.6**

[a] Experimental proton affinities obtained from the NIST chemistry webbook.[ref 23 of Chapter 6] [b] G3MP2 proton affinities were obtained from Gronert et al.[ref 21 of Chapter 6] [c] Anionic proton affinities calculated with CBS-QB3. All errors are computed as $PA^{calc} - PA^{exp}$.

Table B5: Absolute interaction energies of hydrogen bonded base pairs obtained from geometry optimization calculations (kcal/mol)

	Reference	Error							
H-bonded complexes	CCSD(T)[a]	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
2-aminoA...T	-19.5	1.47	10.30	12.62	-277.56	9.77	11.01	9.19	3.28
2-aminoA...T_pl	-19.7	1.42	9.62	11.25	-267.94	7.47	9.78	-35.13	1.24
8oxoG...C_WC	-33.3	2.36	23.86	15.67	139.57	15.01	17.66	73.34	2.97
8oxoG...G	-22.8	1.66	15.19	9.52	189.39	4.28	15.60	140.60	8.56
A...A1_pl	-14.5	1.84	12.56	11.00	636.32	12.68	13.12	12.72	-5.34
A...A2_pl	-13.7	2.05	12.01	15.71	641.81	13.33	14.22	13.05	2.10
A...A3_pl	-12.2	2.14	10.89	6.94	629.45	11.23	13.91	-74.79	-3.15
A...C_pl	-17.6	1.73	13.19	10.19	134.98	14.62	7.24	-52.51	0.61
A...T_WC	-16.86	-0.15	10.67	12.56	-285.25	6.12	6.94	97.92	-3.80
C...C_H+	-51.4	1.6	27.54	21.51	574.50	16.92	15.67	44.10	3.14
G...A1	-19.4	1.5	12.22	7.45	638.59	9.97	11.20	8.70	3.47
G...A1_pl	-18.9	1.64	13.71	11.74	631.59	6.98	11.10	15.77	0.43
G...A2	-14.4	2.74	13.53	9.17	637.81	12.42	8.52	10.65	1.41
G...A2_pl	-12.8	2.05	10.08	5.71	625.19	7.80	0.78	4.90	-1.79
G...A3	-18.8	1.79	14.30	8.71	637.61	9.71	13.19	15.95	2.46
G...A4	-13.5	1.77	9.24	5.67	639.94	11.51	5.71	10.61	-1.62
G...A_HB	-11.3	-0.05	29.27	25.51	673.68	16.55	7.76	-87.30	3.00
G...C_WC	-32.06	1.54	23.83	18.88	141.52	21.39	14.71	67.36	3.25
G...G_pl	-21.3	1.18	11.12	12.39	200.00	11.44	6.60	-98.06	6.98
G...U	-19.1	1.57	12.67	18.03	-424.80	9.49	12.22	66.07	7.89
mA...mT_H	-18.16	2.1	13.19	11.47	-112.65	12.44	10.53	24.18	0.04
mG...mC_WC	-31.59	2.06	20.10	17.27	299.54	4.35	17.60	-158.96	9.23
U...U_calcutta	-10.3	1.04	9.64	6.15	-435.40	0.73	8.33	1.55	5.51
U...U_pl	-13.7	1.22	13.72	8.32	-429.46	0.42	9.86	27.40	5.73
MSE (vs CCSD(T))		**1.59**	**14.69**	**12.23**	**243.27**	**10.28**	**10.97**	**5.72**	**2.32**
MUE (vs CCSD(T))		**1.61**	**14.69**	**12.23**	**419.26**	**10.28**	**10.97**	**47.95**	**3.63**

[a] Energies obtained from ref 27 of Chapter 6. [b] Energies obtained from ref 28 of Chapter 6. All errors are computed as $INT^{calc}–INT^{ref}$.

Table B6: Absolute interaction energies of stacked base pairs obtained from geometry optimization calculations (kcal/mol)

	Reference	Error							
Stacked complexes	CCSD(T)[a]	DFT[b]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
A...C_S	-6.7	1.5	28.75	16.85	164.43	-142.72	22.71	40.02	5.47
A...G_S	-6.5	0.72	16.02	21.69	20.14	16.85	13.85	9.90	6.04
A...T_S2	-8.1	1.32	15.07	5.11	164.38	-156.41	13.32	29.23	-2.67
A...T_S	-12.3	-0.04	6.46	14.95	2.64	15.51	12.82	9.21	3.07
CC_1	2.45	1.33	3.60	2.10	13.72	14.05	-0.48	7.31	2.70
CC_2	-3.85	0.71	7.00	3.05	14.28	14.33	4.16	8.76	4.10
CC_3	-8.88	0.66	8.75	1.93	11.12	14.99	1.91	8.67	-3.95
CC_4	-9.92	0.62	9.41	5.16	11.66	17.58	4.80	9.19	3.71
CC_5	0.32	0.8	2.32	-1.46	11.39	13.45	-1.86	8.49	4.07
CC_6	0.64	0.72	1.65	2.50	11.49	14.46	3.61	8.93	1.34
CC_7	-0.98	0.95	1.74	-2.48	7.48	11.15	-1.01	6.94	1.52
CC_8	-9.1	0.83	8.16	4.93	7.48	15.65	4.69	9.13	1.97
CC_9	-9.11	0.54	7.46	5.26	9.61	16.25	5.09	7.87	0.92
CC_10	-8.27	0.86	7.57	5.76	8.33	16.67	3.91	7.05	-1.37
CC_11	-9.43	0.94	8.96	3.04	8.32	16.69	3.97	8.19	0.83
CC_12	-7.43	1.2	5.71	2.67	2.13	11.64	0.94	7.52	-3.57
CC_13	-8.8	0.58	7.52	4.60	11.72	17.40	4.44	7.25	-1.49
CC_14	-9.11	0.74	8.74	5.36	8.04	16.32	4.73	8.05	-2.46
C...G_S	-12.4	2.3	14.71	22.73	177.89	21.96	17.02	5.06	5.48
G...C_S1	-7.9	2.09	14.43	5.83	-2.70	-159.05	12.77	27.10	-5.13
G...C_S2	-7.7	1.6	16.14	18.52	6.41	12.33	11.54	1.41	14.00
G...C_S3	-11.6	1.42	18.22	21.95	-0.12	17.54	17.00	4.46	5.63
G...C_S	-19.02	-0.24	16.92	18.79	208.20	17.57	15.40	19.08	4.95
mA...mT_S	-14.57	-0.32	11.77	23.77	9.80	18.34	16.27	17.29	9.79
mG...mC_S	-20.35	0.4	12.27	13.47	-182.08	10.15	15.72	10.20	3.26
T...G_S	-6.2	1.95	15.26	11.14	242.72	-142.87	14.37	32.15	-1.22
MSE (vs CCSD(T))		**0.93**	**10.56**	**9.12**	**36.48**	**-10.01**	**8.53**	**12.25**	**2.19**
MUE (vs CCSD(T))		**0.98**	**10.56**	**9.43**	**50.70**	**36.23**	**8.78**	**12.25**	**3.87**

[a] Energies obtained from ref 27 of Chapter 6. [b] Energies obtained from ref 28 of Chapter 6. All errors are computed as INT^{calc}–INT^{ref}.

Table B7: Absolute interaction energies of hydrogen bonded base pairs obtained from single point calculations (kcal/mol)

	Reference	Error						
H-bonded complexes	CCSD(T)[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
2-aminoA...T	-19.5	13.99	11.65	-224.48	10.75	12.33	11.40	4.45
2-aminoA...T_pl	-19.7	13.87	13.22	-222.64	10.93	10.34	9.78	5.10
8oxoG...C_WC	-33.3	21.54	14.27	178.76	9.52	16.65	14.79	6.85
8oxoG...G	-22.8	19.69	10.40	255.20	4.98	17.23	16.85	7.52
A...A1_pl	-14.5	12.63	8.43	681.77	11.55	8.17	11.70	3.15
A...A2_pl	-13.7	13.54	8.02	681.18	11.59	9.31	12.55	4.22
A...A3_pl	-12.2	13.59	8.06	680.83	11.82	10.20	12.76	5.20
A...C_pl	-17.6	13.02	7.40	172.69	10.28	7.49	12.47	2.33
A...T_WC	-16.86	12.50	9.22	-230.60	8.91	10.06	11.06	4.20
C...C_H+	-51.4	26.96	15.84	612.44	11.48	16.01	18.08	3.49
G...A1	-19.4	14.74	9.36	682.22	9.09	11.79	12.76	4.46
G...A1_pl	-18.9	14.15	11.15	685.05	9.62	9.54	10.68	4.96
G...A2	-14.4	7.67	3.74	671.48	8.51	4.03	6.64	4.50
G...A2_pl	-12.8	12.08	7.52	680.25	9.95	5.70	11.29	2.82
G...A3	-18.8	15.19	8.18	680.27	8.81	12.79	13.56	5.07
G...A4	-13.5	11.01	6.68	679.62	9.41	6.13	11.50	1.97
G...A_HB	-11.3	23.09	13.90	686.85	16.53	13.22	20.19	5.11
G...C_WC	-32.06	19.28	15.56	178.45	10.68	15.57	12.74	8.32
G...G_pl	-21.3	10.15	11.48	251.57	9.24	3.98	5.65	3.37
G...U	-19.1	16.80	12.80	-394.10	5.38	11.67	10.68	6.53
mA...mT_H	-18.16	14.77	10.04	-53.81	9.56	10.26	11.64	3.54
mG...mC_WC	-31.59	20.88	14.48	356.58	10.31	16.44	14.55	7.13
U...U_calcutta	-10.3	7.23	6.23	-401.34	2.54	8.88	8.00	5.81
U...U_pl	-13.7	12.19	8.01	-397.27	2.15	11.20	10.32	5.54
MSE (vs CCSD(T))		**15.02**	**10.23**	**287.12**	**9.32**	**10.79**	**12.15**	**4.82**
MUE (vs CCSD(T))		**15.02**	**10.23**	**447.48**	**9.32**	**10.79**	**12.15**	**4.82**

[a] Energies obtained from ref 27 of Chapter 6. All errors are computed as $INT^{calc} - INT^{ref}$.

Table B8: Absolute interaction energies of stacked base pairs obtained from single point calculations (kcal/mol)

	Reference	Error						
Stacked complexes	CCSD(T)[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
A...C_S	-6.7	41.51	26.28	194.59	-135.94	30.99	49.97	10.97
A...G_S	-6.5	21.18	21.97	20.38	18.73	17.24	17.34	11.52
A...T_S2	-8.1	27.91	13.37	183.06	-150.81	19.11	38.73	-2.71
A...T_S	-12.3	15.32	17.25	24.09	16.51	16.23	14.61	10.22
CC_1	2.45	0.85	1.78	4.57	5.26	0.00	0.97	3.86
CC_2	-3.85	3.91	4.33	3.66	5.77	2.48	2.54	5.17
CC_3	-8.88	5.27	5.32	1.42	6.33	3.63	2.17	5.34
CC_4	-9.92	5.76	6.81	2.94	9.04	4.13	2.33	5.67
CC_5	0.32	2.71	3.81	5.80	6.84	1.66	2.66	5.11
CC_6	0.64	2.69	4.13	6.94	7.26	1.71	2.60	5.33
CC_7	-0.98	0.50	-0.23	-2.07	2.04	-0.30	0.46	3.03
CC_8	-9.1	4.54	4.05	-2.08	6.63	3.03	2.12	3.83
CC_9	-9.11	5.31	6.66	4.28	8.48	3.66	1.45	6.25
CC_10	-8.27	4.23	5.58	2.08	8.05	3.12	0.71	4.48
CC_11	-9.43	4.79	5.56	1.74	8.01	2.99	1.52	4.94
CC_12	-7.43	1.35	-0.16	-8.19	1.94	0.04	0.50	1.24
CC_13	-8.8	4.21	5.31	2.48	8.09	3.41	0.75	5.41
CC_14	-9.11	5.15	6.60	2.82	8.84	3.67	1.59	4.99
C...G_S	-12.4	19.35	18.92	13.43	14.48	14.19	13.54	11.08
G...C_S1	-7.9	34.23	20.04	182.87	-145.55	22.64	41.25	2.85
G...C_S2	-7.7	17.79	19.39	18.36	16.69	13.22	12.63	11.04
G...C_S3	-11.6	17.35	17.28	13.08	13.56	13.02	11.26	10.42
G...C_S	-19.02	13.69	17.68	23.34	16.89	17.31	14.24	8.51
mA...mT_S	-14.57	18.50	24.65	33.69	21.80	19.76	17.07	13.05
mG...mC_S	-20.35	18.17	22.79	30.43	21.86	21.71	18.65	9.63
T...G_S	-6.2	33.86	23.72	193.86	-141.39	22.62	39.25	1.99
MSE (vs CCSD(T))		**12.70**	**11.65**	**36.83**	**-13.10**	**10.05**	**11.96**	**6.28**
MUE (vs CCSD(T))		**12.70**	**11.68**	**37.78**	**31.03**	**10.07**	**11.96**	**6.49**

[a] Energies obtained from ref 27 of Chapter 6. All errors are computed as $INT^{calc}–INT^{ref}$.

Table B9: Reference and calculated (single point) interaction energies for a number of stacked carbohydrate-aromatic systems (kcal/mol)

	Reference	Error						
Dimer	DFT[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-Galactose−benzene	-5.14	5.75	4.38	7.44	3.80	4.43	1.87	4.57
β-Glucose−p-hydroxy toluene	-7.27	6.95	5.60	10.67	5.12	6.14	2.40	6.42
α-methyl-glucose−toluene	-7.97	7.80	5.91	11.23	5.98	6.90	4.16	7.06
Fucose−toluene[b]	-6.52	5.54	3.92	10.54	3.82	5.45	1.25	5.75
Fucose−toluene[c]	-9.28	8.64	7.30	12.60	7.04	7.97	4.74	8.40
MSE		**6.94**	**5.42**	**10.49**	**5.15**	**6.18**	**2.89**	**6.44**
MUE		**6.94**	**5.42**	**10.49**	**5.15**	**6.18**	**2.89**	**6.44**

[a] DFT structures and energies were obtained with the DFT-D/TZVDZ level of theory and basis set from work by Raju et al.[ref 29-30 of Chapter 6] [b] Conformer 5(II) of Raju et al. [ref 29-30 of Chapter 6] [c] Conformer 8(I) of Raju et al. [ref 29-30 of Chapter 6] All errors are computed as $INT^{calc} - INT^{ref}$.

Table B10: Reference and calculated (geometry optimized) interaction energies for a number of stacked carbohydrate-aromatic systems (kcal/mol)

	Reference	Error						
Dimer	DFT[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
β-Galactose−benzene	-5.14	6.67	5.20	0.35	4.40	6.87	5.48	6.16
β-Glucose−p-hydroxytoluene	-7.27	5.64	5.00	-10.30	-1.10	5.13	-1.95	1.06
α-methyl-glucose−toluene	-7.97	6.38	5.12	-35.45	7.62	8.36	3.21	7.85
Fucose−toluene[b]	-6.52	6.53	5.23	5.14	5.13	2.72	5.27	6.82
Fucose−toluene[c]	-9.28	9.25	7.51	11.47	7.54	6.55	7.89	10.31
MSE		**6.89**	**5.61**	**-5.76**	**4.72**	**5.93**	**3.98**	**6.44**
MUE		**6.89**	**5.61**	**12.54**	**5.16**	**5.93**	**4.76**	**6.44**

[a] DFT structures and energies were obtained with the DFT-D/TZVDZ level of theory and basis set from work by Raju et al.[ref 29-30 of Chapter 6] [b] Conformer 5(II) of Raju et al. [ref 29-30 of Chapter 6] [c] Conformer 8(I) of Raju et al. [ref 29-30 of Chapter 6] All errors are computed as $INT^{calc} - INT^{ref}$.

Table B11: Barrier heights obtained from QM/MM simulations with SE and high-level quantum mechanical methods

Reaction	State[a]	Reference	Error						
		DFT[c]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
Catalyzed O-GlcNAc transferase[d]	TS	19.6	45.9	38.6	33.5	54.4	34.9	29.1	15.0
	P	-5.8	62.5	54.1	30.4	58.0	43.5	-2.1	19.0
Catalyzed O-GlcNAc transferase[e]	TS	19.6	47.2	38.9	3.7	76.4	31.6	28.8	7.5
	P	-5.8	61.7	55.3	36.0	58.6	38.5	24.3	11.6

[a] Different reaction states, TS = transition state, and P = Product. [c] QM(MPW1K)/MM energy obtained from ref 31 of Chapter 6. [d] SE energies obtained via single point QM/MM calculations on the MPW1K geometries. [e] SE energies obtained via geometry optimized QM/MM calculations. All energy barriers are computed relative to the reactant state. All errors are computed as $\Delta E^{calc} - \Delta E^{ref}$.

Table B12: RMSD of 6-membered ring contained within GT reaction (Å)

Reaction	State[a]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
Catalyzed O-GlcNAc transferase	R	0.012345	0.020410	0.046910	0.049712	0.020833	0.025402	0.013110
	TS	0.026978	0.025342	0.050082	0.058153	0.026842	0.026248	0.018694
	P	0.011491	0.018862	0.041360	0.043504	0.023040	0.026329	0.012373

[a] Different reaction states, R= Reactant, TS = Transition state, and P = Product.

Table B13: Difference in reaction coordinate bond lengths of GT reaction obtained for optimized QM/MM structures (Å)

Reaction	State[a]	RC[b]	Reference	Difference						
			MPW1K[c]	AM1	PM3	PM3CARB-1	$PM3^{MS}$	RM1	AM1/d-PhoT	AM1/d-CB1
Catalyzed O-GlcNAc transferase	R	O_1'-C_1'	1.44	-0.09	-0.12	-0.13	-0.14	-0.11	-0.01	-0.05
		OG-C_1'	3.00	0.05	0.07	0.05	0.09	0.06	0.02	0.05
	TS	O_1'-C_1'	3.11	-0.02	-0.04	0	0.01	-0.06	-0.01	-0.05
		OG-C_1'	1.92	0.04	0.06	0	-0.01	0.07	0	0.08
	P	O_1'-C_1'	3.35	-0.02	-0.04	0.05	0.04	-0.03	-0.04	0.03
		OG-C_1'	1.39	0.04	0.04	-0.01	0.03	0.03	0.06	-0.01

[a] Different reaction states, R= Reactant, TS = Transition state, and P = Product. [b] RC is the reaction coordinate. [c] QM(MPW1K)/MM energy obtained from ref 31 of Chapter 6.

Appendix C

Table C1. Molecular classes used during the parameterization

Class	Molecules
Phosphate	HPO_3, HPO_4^{2-}, $H_2PO_4^-$, H_3PO_4, $(OCH_3)(OH)_2PO$, $(OCH_3)(OH)(O)PO^-$, $(OCH_3)_2(OH)PO$
β-gluc−phos	$^1C_4$-HPO_4^-, $^4C_1$-HPO_4^-, $^1C_4$-H_2PO_4, $^4C_1$-H_2PO_4
α-gluc−phos	$^1C_4$-HPO_4^-, $^4C_1$-HPO_4^-, $^1C_4$-H_2PO_4, $^4C_1$-H_2PO_4
β-ribo−phos	6-HPO_4^-, 12-HPO_4^-, 6-H_2PO_4, 12-H_2PO_4
α-ribo−phos	2-HPO_4^-, 4-HPO_4^-, 2-H_2PO_4, 4-H_2PO_4

Table C2. Experimental and Theoretical gas phase proton affinities for molecules used in parameterization (kcal/mol)

	Reference	Error			
Molecule	DFT[a]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
HPO_3	311.0	-26.9	-2.6	-0.5	-3.0
HPO_4^{2-}	584.9	-11.0	7.9	5.2	11.8
$H_2PO_4^-$	460.0	-26.4	-3.8	-4.9	0.2
H_3PO_4	329.9	-35.4	-8.3	-4.9	-1.2
$(OCH_3)(OH)_2PO$	331.1	-35.9	-11.5	-3.4	-4.2
$(OCH_3)(OH)(O)PO^-$	455.3	-28.9	-6.5	-1.6	-1.5
MUE (vs DFT)		**27.4**	**6.8**	**3.4**	**3.7**
MSE (vs DFT)		**-27.4**	**-4.2**	**-1.7**	**0.4**

[a] The DFT proton affinities where computed with M06-2X/6-311++G(3df,2p) level of theory and basis set. All errors are computed as $PA^{calc} - PA^{ref}$.

Table C3. Relative gas phase proton affinities for selected molecules used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose-phosphate					
$^1C_4$−HPO_4^-	0	4.18	0	0	0
$^4C_1$−HPO_4^-	12.17	0	21.25	8.49	25.84
$^1C_4$−H_2PO_4	0	0	0	0	0
$^4C_1$−H_2PO_4	4.58	7.20	10.39	6.11	4.93
MUE (vs DFT)		**4.74**	**3.72**	**0.54**	**3.51**
MSE (vs DFT)		**-1.34**	**3.72**	**-0.54**	**3.51**
β-D-ribofuranose-phosphate[a]					
6−HPO_4^-	3.09	0	0.89	0.83	3.95
12−HPO_4^-	0	6.25	0	0	0
6−H_2PO_4	0.23	1.83	0.45	4.25	2.15
12−H_2PO_4	0	0	0	0	0
MUE (vs DFT)		**2.73**	**0.60**	**1.57**	**0.69**
MSE (vs DFT)		**1.19**	**-0.50**	**0.44**	**0.69**
α-D-glucopyranose-phosphate					
$^1C_4$−HPO_4^-	0	0	0	0	0
$^4C_1$−HPO_4^-	1.56	0.02	2.46	7.08	2.71
$^1C_4$−H_2PO_4	0	5.63	0	0	0
$^4C_1$−H_2PO_4	4.34	0	12.64	0.82	12.84
MUE (vs DFT)		**2.88**	**2.30**	**2.26**	**2.41**
MSE (vs DFT)		**-0.06**	**2.30**	**0.50**	**2.41**
α-D-ribofuranose-phosphate					
2−HPO_4^-	5.09	0	4.60	0.91	0
4−HPO_4^-	0	6.77	0	0	0.29
2−H_2PO_4	0.40	1.89	1.44	0.59	0
4−H_2PO_4	0	0	0	0	0.45
MUE (vs DFT)		**3.34**	**0.38**	**1.09**	**1.56**
MSE (vs DFT)		**0.79**	**0.14**	**-1.00**	**-1.19**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT proton affinities obtained with M06-2X/6-311++G(3df,2p).

Table C4. Absolute dipole moments of selected molecules used in parameterization (Debye)

	Reference	Error			
Molecule	DFT[b]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
HPO_3	3.34	-0.11	-0.24	0.16	-0.05
$P(CH_3)_3$	1.22	-0.02	0.28	0.91	3.21
$(CH_3)_3P(O)$	4.47	-0.53	-0.09	0.84	1.20
H_3PO_4	0.49	-0.34	-0.37	-0.43	-0.32
$(OCH_3)(OH)_2PO$	0.94	0.04	-0.09	0.19	0.07
$(OCH_3)_2(OH)PO$	1.14	0.00	-0.15	0.13	0.06
$(OCH_3)_3PO$	1.05	-0.40	-0.47	-0.61	-0.27
MUE (vs DFT)		**0.21**	**0.24**	**0.47**	**0.74**
MSE (vs DFT)		**-0.19**	**-0.16**	**0.17**	**0.56**
β-D-glucopyranose-phosphate					
$^1C_4{-}H_2PO_3$	10.0	-0.44	-0.09	1.35	0.79
$^4C_1{-}H_2PO_3$	4.85	0.32	0.02	0.74	0.54
α-D-glucopyranose-phosphate					
$^1C_4{-}H_2PO_3$	5.16	0.24	0.16	0.94	0.68
$^4C_1{-}H_2PO_3$	3.73	0.24	0.24	0.76	0.74
MUE (vs DFT)		**0.31**	**0.13**	**0.95**	**0.69**
MSE (vs DFT)		**0.09**	**0.08**	**0.95**	**0.69**
β-D-ribofuranose-phosphate[a]					
$6{-}H_2PO_3$	3.97	0.29	0.004	0.44	0.44
$12{-}H_2PO_3$	3.76	0.25	-0.07	0.24	0.36
α-D-ribofuranose-phosphate					
$2{-}H_2PO_3$	6.52	0.24	0.21	1.56	0.91
$4{-}H_2PO_3$	6.64	0.18	-0.08	0.99	0.58
MUE (vs DFT)		**0.24**	**0.09**	**0.81**	**0.57**
MSE (vs DFT)		**0.24**	**0.02**	**0.81**	**0.57**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT dipole moments obtained with M06-2X/6-311++G(3df,2p). All error are computed as $\mu^{calc} - \mu^{DFT}$.

Table C5. Absolute ionization potentials of selected molecules used in parameterization (eV)

	Reference	Error			
Molecule	DFT[b]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
$P(CH_3)_3$	8.56	1.43	0.43	-0.90	0.36
$(CH_3)_3P(O)$	9.85	0.95	0.64	0.01	0.76
H_3PO_4	11.63	2.25	0.82	0.38	-0.35
HPO_3	12.71	1.86	1.02	0.12	0.56
$(OCH_3)(OH)_2PO$	11.29	1.51	1.00	0.33	0.69
$(OCH_3)_2(OH)PO$	11.06	1.57	1.10	0.34	0.76
$(OCH_3)_3PO$	10.90	1.60	1.11	0.32	0.79
MUE (vs DFT)		**1.60**	**0.87**	**0.34**	**0.61**
MSE (vs DFT)		**1.60**	**0.87**	**0.09**	**0.51**
β-D-glucopyranose-phosphate					
$^1C_4-H_2PO_3$	10.38	0.86	0.97	0.59	1.09
$^4C_1-H_2PO_3$	10.46	0.95	1.08	0.67	1.17
α-D-glucopyranose-phosphate					
$^1C_4-H_2PO_3$	10.01	0.89	1.05	0.72	1.22
$^4C_1-H_2PO_3$	10.54	0.98	1.06	0.72	1.23
MUE (vs DFT)		**0.92**	**1.04**	**0.68**	**1.18**
MSE (vs DFT)		**0.92**	**1.04**	**0.68**	**1.18**
β-D-ribofuranose-phosphate[a]					
$6-H_2PO_3$	10.59	0.85	0.89	0.67	1.02
$12-H_2PO_3$	10.60	0.83	0.88	0.65	1.02
α-D-ribofuranose-phosphate					
$2-H_2PO_3$	10.48	0.77	0.79	0.62	0.95
$4-H_2PO_3$	10.80	1.02	1.07	0.80	1.08
MUE (vs DFT)		**0.87**	**0.91**	**0.69**	**1.02**
MSE (vs DFT)		**0.87**	**0.91**	**0.69**	**1.02**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5]. [b] DFT ionization obtained with M06-2X/6-311++G(3df,2p). All errors are computed as $IP^{calc} - IP^{DFT}$.

Table C6. Experimental and Theoretical Heats of formation for molecules used in parameterization (kcal/mol)

	Reference		Error			
Molecule	Exp	DFT[c]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
$P(CH_3)_3$	-24.2[a]	-24.1	8.9	-12.1	16.0	-11.1
$(CH_3)_3PO$	-103.8[a]	-95.7	0	17.1	4.4	1.9
PO_3^-	-225.4[a]	-221.3	-37.8	-2.3	19.7	10.9
H_3PO_4	-272.8[b]	-272.2	11.8	21.7	34.8	14.6
MUE (vs exp)		**3.2**	**14.6**	**13.3**	**18.7**	**9.6**
MSE (vs exp)		**3.2**	**-4.3**	**6.1**	**18.7**	**4.1**

[a] Experimental values obtained from Nam et al.[ref 8 of Chapter 5] [b] Theoretical values derived by Alexeev et al.[ref 17 of Chapter 5] [c] The DFT heats of formation where computed with M06-2X/6-311++G(3df,2p) level of theory and basis set. All errors are computed as $\Delta H_{\text{int}}^{calc} - \Delta H_{\text{int}}^{\text{exp}}$.

Table C7. Relative heats of formation for different protonated forms of carbohydrate-phosphate conformers used in parameterization (kcal/mol)

Molecule	DFT[b]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
β-D-glucopyranose-phosphate					
$^1C_4$−H_2PO_4	4.20	3.58	9.04	0	0.56
$^4C_1$−H_2PO_4	0	0	0	9.33	0
$^1C_4$−HPO_4^-	0	0	0	0	0
$^4C_1$−HPO_4^-	0.02	3.61	1.35	15.44	4.37
$^1C_4$−PO_4^{2-}	0	1.00	0	0	0
$^4C_1$−PO_4^{2-}	12.19	0	22.59	23.93	30.21
MUE (vs DFT)		**2.83**	**2.76**	**6.78**	**4.33**
MSE (vs DFT)		**-1.44**	**2.76**	**5.38**	**3.12**
β-D-ribofuranose-phosphate[a]					
6−H_2PO_4	0	3.80	2.52	2.10	0
12−H_2PO_4	1.77	0	0	0	0.49
6−HPO_4^-	0	5.63	2.97	6.35	1.66
12−HPO_4^-	1.64	0	0	0	0
6−PO_4^{2-}	1.54	0	3.86	7.18	5.61
12−PO_4^{2-}	0	0.62	0	0	0
MUE (vs DFT)		**2.50**	**1.87**	**2.92**	**1.44**
MSE (vs DFT)		**0.85**	**0.73**	**1.78**	**0.47**
α-D-glucopyranose-phosphate					
$^1C_4$−H_2PO_4	8.11	5.11	5.50	4.68	9.66
$^4C_1$−H_2PO_4	0	0	0	0	0
$^1C_4$−HPO_4^-	3.73	10.74	0	3.86	0
$^4C_1$−HPO_4^-	0	0	7.14	0	3.18
$^1C_4$−PO_4^{2-}	2.11	10.72	0	0	0
$^4C_1$−PO_4^{2-}	0	0	9.60	3.22	5,89
MUE (vs DFT)		**3.10**	**4.20**	**1.48**	**2.74**
MSE (vs DFT)		**2.10**	**1.38**	**-0.37**	**0.80**
α-D-ribofuranose-phosphate					
2−H_2PO_4	2.92	0	0	5.21	4.66
4−H_2PO_4	0	2.38	0.83	0	0
2−HPO_4^-	3.14	0	0.61	5.81	4.22
4−HPO_4^-	0	0.48	0	0	0
2−PO_4^{2-}	8.18	0	5.21	6.72	3.93
4−PO_4^{2-}	0	7.26	0	0	0
MUE (vs DFT)		**4.06**	**1.54**	**1.07**	**1.18**
MSE (vs DFT)		**-0.69**	**-1.27**	**0.58**	**-0.24**

[a] Initial ring conformations (without the phosphate) obtained from Jalbout et al.[ref 15 of Chapter 5] [b] DFT energies obtained with M06-2X/6-311++G(3df,2p).

Table C8: Barrier heights obtained from QM/MM simulations with SE and high-level quantum mechanical methods

Reaction	State[a]	Reference	Error			
		DFT[c]	AM1*	AM1*-CB1	AM1/d-PhoT	AM1/d-CB1
Catalyzed O-GlcNAc transferase[d]	TS	19.6	13.0	-2.7	29.1	15.0
	P	-5.8	31.9	3.2	-2.1	19.0
Catalyzed O-GlcNAc transferase[e]	TS	19.6	15.8	-6.8	28.8	7.5
	P	-5.8	31.3	0.4	24.3	11.6

[a] Different reaction states, TS = transition state, and P = Product. [c] QM(MPW1K)/MM energy obtained from ref 31 of Chapter 6. [d] SE energies obtained via single point QM/MM calculations on the MPW1K geometries. [e] SE energies obtained via geometry optimized QM/MM calculations. All energy barriers are computed relative to the reactant state. All errors are computed as $\Delta E^{calc} - \Delta E^{ref}$.