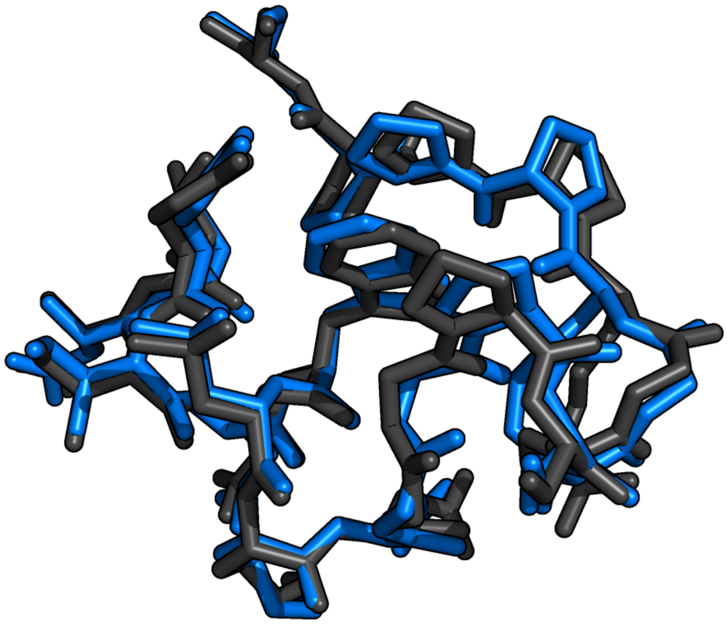
*Jimmy Charnley Kromann*

# New Methods for Computational Drug Design

MASTER'S THESIS



*Department of Chemistry · University of Copenhagen*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.799v1 | CC-BY 4.0 Open Access | rec: 16 Jan 2015, publ: 16 Jan 2015

# Abstract

This thesis describes the work that has been carried out in connection with my Masters at the University of Copenhagen. This work has led to new dispersion and hydrogen bond corrections to the PM6 method, PM6-D3H+, and its implementation in the GAMESS program. The method combines the DFT-D3 dispersion correction by Grimme *et al.* with a modified version of the H+ hydrogen bond correction by Korth. This work also included the implementation of the new HF-3c method in GAMESS and its interface with the fragmentation method FMO.

Overall, the interaction energy of PM6-D3H+ is very similar to PM6-DH2 and PM6-DH+, with RMSD and MAD values within 0.02 kcal/mol of one another. HF-3c also shows interaction energies within the same order of accuracy as the PM6 based methods. The main difference is that the geometry optimizations of 88 complexes result in 82, 6, 0, and 0 geometries with 0, 1, 2, and 3 or more imaginary frequencies using PM6-D3H+ implemented in GAMESS, while the corresponding numbers for PM6-DH+ implemented in MOPAC are 54, 17, 15, and 2. PM6-D3H+ and FMO2-HF-3c in GAMESS was used to optimize two small proteins which resulted in a much more reliable structure compared to the reference structures, than PM6-DH+ in MOPAC, most likely due to the different optimization algorithms associated with the programs.

The PM6-D3H+ method as implemented in GAMESS offers an attractive alternative to PM6-DH+ in MOPAC in cases where the LBFGS optimizer must be used and a vibrational analysis is needed, e.g., when computing vibrational free energies. While the GAMESS implementation is up to 10 times slower for geometry optimizations of proteins in bulk solvent compared to MOPAC, it is sufficiently fast to make geometry optimizations of small proteins practically feasible.

# Acknowledgements

This thesis represents the work I have carried out as a Master's student under the supervision of Professor Jan Halborg Jensen in his group of Bio-computational Chemistry at the Department of Chemistry, University of Copenhagen. I would especially like to thank the following people:

- Thank you, Jan Jensen, for patience in supervising me and giving me the freedom to make my templates.
- Thank you, X-Stream® people of C315b, Lars A. Bratholm, Casper Steinmann, Maher Channir, Mads B. Færch-Jensen and the lovely Anders S. Christensen, for useful and useless discussions in the office.
- Thank you, Kurt Mikkelsen, Stephan Sauer and Sten rettrup for teaching me quantum chemistry.
- Thank you, the collaborators Hari Muddana and Mike Gilson from UC San Diego and especially Martin Korth from University of Ulm.
- Thank you, Kasper P. and Anders B., for no personal encouragement and total lack of support.
- Thank you, the pharma lunch group, Beyer, Elm, Lykkebo, A. J., Qian, Alessandro, Luca, Falco and Strange.
- Thank you, Google, for answering all my questions.

Lastly, my family for patience in my pursuit of a nearly non-profit science career, instead of getting a real job.

# Contents

# Chapter 1

# Introduction

The ability to produce accurate quantum mechanical (QM) predictions for large-scale bio-molecular systems has the potential to bring new insight into several scientific fields such as enzyme-design and protein-ligand docking. The field is but still confronted with serious challenges. We can get numerically exact solutions with high-level correlated methods (such as CCSD(T)/CBS)[1], but this is currently too computationally costly for any real bio-chemical relevant systems, such as enzymes. Using classical force-fields one can get very fast results, however, force-fields lack the proper treatment of electron behavior and therefore cannot simulate most chemical reaction.

If you want to move quantum chemistry into bio-chemistry you will need to focus on fast methods, which involves either the standard Hartree-Fock (HF) approach or the semi-empirical quantum methods (SQM) formalism introduced by John Pople, such as NDDO and later PM6[2]. For bio-chemical systems non-covalent interactions are the key to getting good results, however both the HF and PM6 approach, with their many approximations, are in most cases incapable of simulating such interaction, and purely empirical correction terms are introduced.

Dispersion and hydrogen bonded corrections to the PM6 method such as PM6-DH2[3], PM6-D3H4[4] and PM6-DH+[5] yield interaction energies that in many cases rival in accuracy those computed with Density Functional Theory (DFT)[6, 7]. The computational efficiency of the underlying PM6 method allows for calculations that are not practically possible with DFT or HF, such as geometry optimizations of proteins or vibrational analyses of large systems. For example, recent studies by Gilson[8] and Grimme[9] have used dispersion and hydrogen bonded PM6 (PM6-DH+ and PM6-D3H respectively) to compute the vibrational free energy contribution to the standard binding free energy for host-guest systems and have demonstrated that they make a crucial contribution.

However, computing this vibrational free energy contribution can be complicated by the presence of one or more imaginary frequencies in the vibrational analysis[10]. The source of these imaginary frequencies are usually numerical errors amplified by a flat potential energy surface and the imaginary frequencies often correspond to low lying frequencies that make a significant contribution to the vibrational entropy. Thus, these numerical problems can introduce a significant error in the binding free energy. Preliminary calculations suggested that one of the sources of the imaginary frequencies in PM6-DH+ calculations using MOPAC could be solved by using different geometry optimization algorithms.

The work presented in this thesis is a new variant of the MOPAC based method PM6-DH+, called PM6-D3H+[11], in the GAMESS program[12] to allow us to test the use of the optimization algorithms implemented therein. As well as implementation of the newly developed corrected Hatree-Fock model called HF-3c[13], introduced by Sure and Grimme. PM6-D3H+ differs from PM6-DH+ in that the dispersion term is the third generation dispersion model developed by Grimme *et al.*[14] rather than the Jurecka-type model developed by Jurecka *et al.*[15]. In that respect, PM6-D3H+ is identical to the PM6-D3H model developed by Grimme[9] which has not yet been incorporated into a quantum chemistry program. This dispersion model was mainly chosen for convenience (as it was already implemented in GAMESS) and has little effect on the average accuracy compared to PM6-DH+ (although the maximum errors observed for the training set decrease).

The rest of the thesis is concerned with introduction of the underlying methodology for approximation used in these semi-empirical methods, and how these QM methods are so fast. Chapter 2 is a short introduction to the regular electronic problem. Chapter 3 is a introduction to various correction schemes that corrects the electronic models. Chapter 4 presents results of the methods and a discussion of the results obtained using the new SQM methods. Chapter 5 presents a short summary of the work, as well as discussion about the outlook and direction for these type of methods.

# Chapter 2

# Calculating the Electronic Energy

Solving the behavior of electrons in a molecule is a difficult task. However, the alternative, to calculate properties without electrons, such as force fields which purely classical, is not an option for most chemical cases as no electronic treatment leads to no bond-breaking and therefore no chemical reactions. We need a way to describe the behavior of the electrons, if we want to simulate chemical reactions. The purpose of the following derivations is to obtain the theoretical background and the approximations necessary to understand the different approaches, level of accuracy and computational speed of which the electrons of molecules are treated.

## 2.1 The Schrödinger Equation

We can describe the behavior of the molecules by calculating the kinetic and potential energy of the electrons and nuclei. This is done by the Schrödinger equation

$$\hat{\mathcal{H}} \left|\Psi\right\rangle = \mathcal{E}_i \left|\Psi\right\rangle \tag{2.1}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian describing a system of $N$ electrons and $M$ nuclei, $\mathcal{E}$ is the energy of the **stationary** state described by the corresponding wave function $\left|\Phi\right\rangle$. The Hamiltonian for $N$ electrons and $M$ nucleus is defined as

$$\hat{\mathcal{H}} = -\sum_i^N \frac{1}{2}\hat{\nabla}_i^2 - \sum_i^N \sum_A^M \frac{Z_A}{r_{iA}} + \sum_i^N \sum_{j>i}^N \frac{1}{r_{ij}} + \sum_A^M \sum_{B>A}^M \frac{Z_A Z_B}{r_{AB}} - \sum_A^M \frac{1}{2M_A}\hat{\nabla}_A^2 \tag{2.2}$$

where the first term is the kinetic energy operator, with the Laplace operator which is a second order differential operator. The second term is the Coulomb attraction between electrons and nuclei, here $Z_A$ is the atomic charge of nucleus $A$ and $r_{iA}$ the distance $|\mathbf{r}_i - \mathbf{r}_A|$ between electron $i$ and nucleus $A$. The third and fourth term is Coulomb repulsion between electron-electron and nuclei-nuclei, respectively. Here $r_{ij}$ is the distance $|\mathbf{r}_i - \mathbf{r}_j|$ between electron $i$ and electron $j$ and $r_{AB}$ the distance $|\mathbf{r}_A - \mathbf{r}_B|$ between nuclei $A$ and nuclei $B$. The fifth term is the kinetic energy of the nuclei, where $M_A$ is the mass of nuclei $A$.

In the above Hamiltonian the movement of nuclei have been neglected, because we work within the **Born-Oppenheimer** approximation, which states that because electrons are much lighter than the nuclei, the nuclei appears stationary from the electrons perspective. This means we can approximately treat the movement of electrons and nuclei separately. This means we will also leave out the nuclear-nuclear repulsion term (last term) in eq. 2.2, because it can be considered constant. And thus the *electronic* Hamiltonian can be written as

$$\hat{\mathcal{H}}_{\text{elec}} = -\sum_i^N \frac{1}{2}\hat{\nabla}_i^2 - \sum_i^N \sum_A^M \frac{Z_A}{r_{iA}} + \sum_i^N \sum_{j>i}^N \frac{1}{r_{ij}} \tag{2.3}$$

which is the Hamiltonian we use from here on, and thus remove the subscript. This equation for the electrons is impossible to solve analytically for a many-electron system, and thus more approximations are needed.

## 2.2 Hartree-Fock Theory

Solving the Schrördinger equation for a many-electron system is practically impossible, and thus we need to approximate the model. For this we use the Hatree-Fock theory. Almost all *ab initio* computational quantum-chemistry methods are based on **Hartree-Fock** (HF) theory, also called the self-consistent field (SCF) method. In closed-shell HF theory the unperturbed many-electron wave function $|\Psi\rangle$ is approximated by a single anti-symmetric orbital-based wave function, called a Slater determinant.

$$|\Psi\rangle \approx \left|\Phi^{\mathrm{SCF}}\right\rangle \tag{2.4}$$

From the expression of the Slater determinant and the electronic Hamiltonian (eq. 2.3) it is possible, by use of the variation principle and by integrating out the spin functions[16, 17, 18] to arrive at the closed-shell Hartree-Fock equation, also known as **restricted Hartree-Fock** (RHF).

$$\hat{f}(\mathbf{r}_1)\phi_i(\mathbf{r}_1) = \epsilon_i\phi_i(\mathbf{r}_1) \tag{2.5}$$

where $\epsilon_i$ is the energy of the $i$'th orbital $\phi_i$. The problem is now reduced to finding a set of unknown spatial orbitals $\{\phi\}$ under the constraints that the orbitals are orthonormal

$$\langle\phi_p|\phi_q\rangle = \delta_{pq} \tag{2.6}$$

where $\delta_{pq}$ is the Kronecker delta function. The operator $\hat{f}$ is the Fock operator and is given as

$$\hat{f}(\mathbf{r}_1) = \hat{h}(\mathbf{r}_1) + \hat{G}(\mathbf{r}_1) \tag{2.7}$$

Here $\hat{h}$ is the operator of the kinetic energy of an electron plus its attraction to the nuclei (one-electron operator) and $\hat{G}$ is the two electron repulsion (two-electron operator). $\hat{h}$ is the one-electron part of the electronic Hamiltonian (eq. 2.3) for one electron and is given as

$$\hat{h}(\mathbf{r}_1) = -\frac{1}{2}\nabla_1^2 - \sum_k^M \frac{Z_k}{|\mathbf{r}_1 - \mathbf{R}_k|} \tag{2.8}$$

The two electron repulsion part of eq. 2.3 is then described by the operator $\hat{G}$, which is the electron in the mean-field of all the other electrons. The two electron repulsion term is given as

$$\hat{G}(\mathbf{r}_1) = \sum_a^{N/2}\left[2\hat{J}_a(\mathbf{r}_1) - \hat{K}_a(\mathbf{r}_1)\right] \tag{2.9}$$

Here $\hat{J}_a(\mathbf{r}_1)$ and $\hat{K}_a(\mathbf{r}_1)$ are the closed-shell Coulomb and exchange operators, respectively. The exchange operator $\hat{K}$ permutes electron 1 with 2. The total potential operator $\hat{H}$ is the potential that an electron experiences when it moves in the averaged field of all the other electrons. In order to calculate this averaged field of the other electrons one needs molecule orbitals that describe the other electrons. The Fock operator thus depends on its own eigenfunctions and the Hartree-Fock equations have to be solved iteratively until self-consistency of the Hartree-Fock potential is obtained, hence the name self-consistent field method. The two operators operating on a wave function is given as

$$\hat{J}_a(\mathbf{r}_1)\phi_b(\mathbf{r}_1) = \left[\int d\mathbf{r}_2\phi_a^*(\mathbf{r}_2)\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\phi_a(\mathbf{r}_2)\right]\phi_b(\mathbf{r}_1) \tag{2.10}$$

$$\hat{K}_a(\mathbf{r}_1)\phi_b(\mathbf{r}_1) = \left[\int d\mathbf{r}_2\phi_a^*(\mathbf{r}_2)\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\phi_b(\mathbf{r}_2)\right]\phi_a(\mathbf{r}_1) \tag{2.11}$$

where integration is carried out only over spatial coordinates, since we are within the restricted Hartree-Fock approximation. The Coulomb operator $\hat{J}_a$ represents the average local potential at a position $\mathbf{r}_1$ arising from an electron in the orbital $\phi_a$ at position $\mathbf{r}_2$. When one integrates and sums over all the other electrons, the average potential that one electron feels of the other $N-1$

electrons is obtained. The exchange operator, having no classical interpretation, can be seen as the exchange of electron 1 and electron 2 on the right hand side of the eq. 2.11.
If we want to find the orbitals $\{\Phi\}$ we have to do another approximation, introduced by Roothaan, who showed how one could introduce a spatial basis set $\{\chi\}$ and convert the HF equations to be in this known spatial basis space. Using this, we can expand the unknown spatial *molecular* orbitals (MO) in a known basis of *atomic* orbitals (AO). This approximation is know as the **linear combination of atomic orbitals** LCAO.

$$\phi_i = \sum_{\mu}^{K} C_{\mu i}\chi_\mu \quad i = 1, 2, \ldots, K \tag{2.12}$$

where $K$ is the number of AO functions we have chosen to include and $C_{\mu i}$ is the MO coefficient. Following ref. [19, 18] by inserting eq. 2.12 into eq. 2.7 one ends up with the Roothaan equation which is written in a compact form of a single matrix equation

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{2.13}$$

where $\mathbf{F}$ is the Fock matrix, $\mathbf{C}$ the coefficient matrix, $\mathbf{S}$ the overlap matrix and $\epsilon$ the orbital energy matrix in the basis sec $\psi_\mu$. Having expanded everything in the basis $\{\chi_\mu\}$ the individual elements of the Fock matrix are written as

$$F_{\mu\nu} = \langle\chi_\mu|\,\hat{f}(\mathbf{r}_1)\,|\chi_\nu\rangle \tag{2.14}$$

$$= \langle\chi_\mu|\,\hat{h}(\mathbf{r}_1)\,|\chi_\nu\rangle + \sum_{a}^{N/2}\langle\chi_\mu|\,2\hat{J}_a(\mathbf{r}_1) - \hat{K}_a(\mathbf{r}_1)\,|\chi_\nu\rangle \tag{2.15}$$

$$= H_{\mu\nu} + \sum_{a}^{N/2}\sum_{\lambda\sigma}^{K} C_{\lambda a}C_{\sigma a}\left[2\langle\chi_\mu\chi_\nu|\chi_\lambda\chi_\sigma\rangle - \langle\chi_\mu\chi_\sigma|\chi_\lambda\chi_\nu\rangle\right] \tag{2.16}$$

By calculating the appropriate overlap integrals $S_{\mu\nu}$, the core-Hamiltonian integrals $H_{\mu\nu}$ and the two electron integrals it is possible to setup the matrix eq. 2.13 and solve for the coefficient matrix. Because the Fock matrix depends on the coefficient we try to find we need to guess on a set of coefficient and then solve iteratively until the convergence on the density matrix (product of the coefficients) is obtained. Which means the coefficient we get from solving the Fock matrix and the coefficient we insert is the same. The procedure is called the self consistent field (SCF) procedure. In practise the extended Hückel[20] method is often used for the initial guess for the density. The total energy is then the sum of the Hatree-Fock energy and the nuclei-nuclei repulsion term from eq. 2.2

$$E_{\text{total}} = E_{\text{elec}} + \sum_{A}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{r_{AB}} \tag{2.17}$$

## 2.3 The Semi-empirical Way

The most difficult and time-consuming part of LCAO self-consistent molecular orbital calculations is the evaluation and handling of large a number of electron repulsion integrals, and the associated computational cost. For Hartree-Fock calculations the Coulomb and exchange integrals involving molecular orbitals scales in the order of $N^2$ for $N$ orbitals, but the integrals when using the LCAO approximation, scales in the order of $K^4$ for $K$ basis functions, as seen in the elements of the Fock matrix eq. 2.16.

There are different approaches to reduce the amount of integrals we need to evaluate. The first step is of course to keep the number of basis functions $K$ small, essentially all semi-empirical models adapt an AO basis functions that cover only the valence electrons and only has one function per orbital. The basis functions themselves are taken to be Slater-type orbital (STO).

Even though the overlap between any two STO's can be computed analytically (based on the exponents and position) the overlap matrix $S$ for semi-empirical methods are set to a unit matrix.

$$S_{\mu\nu} = \delta_{\mu\nu} \tag{2.18}$$

which means the overlap matrix becomes only non-zero for diagonal terms and simplifies the computation when solving the Roothaan equation (eq. 2.13).

The next approximation is to consider all the electron repulsion integrals needed when building up the Fock matrix, as some of the repulsion integrals have values close to zero. Especially those that includes the overlap $\chi_\mu \chi_\nu$ where $\mu \neq \nu$, or when the basis functions are centered far apart. The most aggressive approximation for neglect Fock matrix integrals is the **zero-differential overlap** (ZDO) approximation, whereby electron repulsion integrals involving the overlap distribution of two different basis functions is assumed negligibly small.

$$\langle \chi_\mu \chi_\nu | \chi_\lambda \chi_\sigma \rangle = \delta_{\mu\nu} \delta_{\lambda\sigma} \langle \chi_\mu \chi_\mu | \chi_\lambda \chi_\lambda \rangle \tag{2.19}$$

where $\delta$ is the Kronecker delta. The core integrals $\hat{H}_{\mu\nu}$ are not neglected but may be treated in a semi-empirical manner to accommodate the possible bonding effect of the overlap. Computationally this brings the integrals down to scale like $K^2$, but at the cost of a lot of integrals and therefore a lot of electron repulsion.

The most elementary theory retaining the main features of electron repulsion is the **complete neglect of differential overlap method** (CNDO). Only valence electrons are treated explicitly, the inner shells being treated as part of a rigid core, so that they modify the nuclear potential in the one-electron part of the Hamiltonian. The atomic orbital basis set $\{\chi_\mu\}$ is then a valence basis set. E.g. $1s$ for Hydrogen and $2s$, $2px$, $2py$, $2pz$ for Carbon. The basic approximation is that the zero-differential overlap approximation is used for all products of different atomic orbitals, however it also includes the additional approximation of making the remaining two-electron integrals depend only on the nature of the atoms A and B to which $\chi_\mu$ and $\chi_\nu$ belong and not on the actual type of orbital.

$$\langle \chi_\mu \chi_\mu | \chi_\lambda \chi_\lambda \rangle = \gamma_{\mathrm{AB}} \begin{cases} \text{all } \mu \text{ on atom A} \\ \text{all } \lambda \text{ on atom B} \end{cases} \tag{2.20}$$

$\gamma_{\mathrm{AB}}$ is then an average electrostatic repulsion between any electron on A and any electron on B. For large distances $r_{\mathrm{AB}}$ between A and B, $\gamma_{\mathrm{AB}}$ will go towards the limit of $r_{\mathrm{AB}}^{-1}$. This approximation does not distinguishing between different electron repulsions. That is to distinguish between different orbitals located on the same atom, as the integral would be the same.

The CNDO approximation introduced electron-electron repulsions in the simplest possible manner. However QM requires that electrons of parallel spin may not occupy the same small region of space and that consequently, two electron in different atomic orbitals on the same atom will have a smaller average repulsion energy if they have parallel spin. Mathematically, this difference shows up as a two-electron exchange integral of the type $\langle \chi_\mu \chi_\nu | \chi_\mu \chi_\nu \rangle$, where $\mu$ and $\nu$ are on the same atom. In CNDO theory such integral are neglected and all interactions between two electrons on atom A are replaced by $\gamma_{\mathrm{AA}}$. Intermediate neglect of differential overlap (INDO) theory fixes this by taking exchange terms into account by retaining mono-atomic differential overlap but only in one-center integrals, which are then parameterized to account for the remaining repulsion.

In **Neglect of Diatomic Differential Overlap** (NDDO)[21] there are a number of additional bicentric integrals to be considered, which involve one-center differential overlap which are neglected in CNDO and INDO. The NDDO approximation is primarily defined as:

$$\langle \chi_\mu^{\mathrm{A}} \chi_\nu^{\mathrm{B}} | \chi_\lambda^{\mathrm{C}} \chi_\sigma^{\mathrm{D}} \rangle = \delta_{\mathrm{AB}} \delta_{\mathrm{CD}} \langle \chi_\mu \chi_\nu | \chi_\lambda \chi_\sigma \rangle \tag{2.21}$$

where the indices A, B, C and D represents atoms in the molecule, which means the overlap matrix is neglected if the two basis functions are centered on different atoms. While a number of semi-empirical methods had been based on the CNDO and INDO formalisms, little attention had been paid to the more rigorous NDDO approximation before the expansion of the integrals in terms of multipole-multipole interactions[21] and the **Modified Neglect of Differential Overlap** MNDO[22, 23] by Walter Thiel and co-workers. The diatomic integrals are not easily evaluated and so instead, the NDDO methods model the integrals as classical multipole interaction between the two atoms. Whether the multipole is a point charge (ss), a dipole (sp) or a quadrupole (pp) depends on the nature of the orbitals. The magnitude of the dipoles and quadrupoles depends on the exponents of the Slater-type basis functions. The non-zero NDDO repulsion integrals thus are either one-center ($\mathrm{A} = \mathrm{B}$) or two-center ($\mathrm{A} \neq \mathrm{B}$) integrals. The two-center integrals represent the electrostatic interactions between the charge distributions at atom A and at atom B. For one-center electron repulsion integrals the same type of values as INDO are retained.

In the next equations we shall assume that the basis functions $\chi_\mu$, and $\chi_\nu$, are centered at atom A, and the basis functions $\chi_\lambda$ and $\chi_\sigma$ at atom B ($\mathrm{A} \neq \mathrm{B}$). In this notation, a diagonal element of the NDDO Fock matrix elements is given by:

$$\begin{aligned} F_{\mu\mu} = U_{\mu\mu} - \sum_{\mathrm{B}\neq\mathrm{A}} Z_{\mathrm{B}} \langle \chi_\mu\chi_\mu | s_{\mathrm{B}} s_{\mathrm{B}}\rangle + \sum_{\nu\in\mathrm{A}} P_{\nu\nu} \left( \langle \chi_\mu\chi_\mu|\chi_\nu\chi_\nu\rangle - \frac{1}{2}\langle\chi_\mu\chi_\nu|\chi_\mu\chi_\nu\rangle \right) \\ + \sum_{\mathrm{B}\neq\mathrm{A}} \sum_{\lambda\in\mathrm{B}} \sum_{\sigma\in\mathrm{B}} P_{\lambda\sigma} \langle\chi_\mu\chi_\mu|\chi_\lambda\chi_\sigma\rangle \end{aligned} \tag{2.22}$$

where $\chi_\mu$ is located on atom A. One-center one-electron energies $U_{\mu\mu}$ represents the sum of the kinetic energy of an electron in $\chi_\mu$ at atom A and its potential energy due to the attraction by the core of atom A, is a atom-specific quantity that is used as an empirical parameter. The second term represents the Coulomb attraction for electrons on atom A and the other nuclei. In the third and fourth term represents the electron-electron repulsion. In the third term, both basis function are on atom A, and the integral overlap are independent of the molecule of interest. There are a fixed set of integrals involving the overlap of the different atomic basis functions, such as $\langle s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{A}} s_{\mathrm{A}}\rangle$, $\langle s_{\mathrm{A}} s_{\mathrm{A}} | p_{\mathrm{A}} p_{\mathrm{A}}\rangle$, and so on. The values of these integrals are obtained by empirical fitting. In the fourth term, only Coulomb integrals remain, and these are further approximated using point charges and classical Coulomb charge repulsion (Coulomb's law).

Similar considerations were made to the off-diagonal Fock matrix element $F_{\mu\lambda}$, where $\mu \neq \lambda$. However, it is also assumed that the overlap of the kinetic energy $\langle\chi_\mu| - \frac{1}{2}\nabla^2 |\chi_\lambda\rangle$ is zero if $\chi_\mu$ and $\chi_\lambda$ are on different atoms (A and B).

$$F_{\mu\lambda} = \frac{1}{2}(\beta_\mu + \beta_\lambda) S_{\mu\lambda} - \frac{1}{2} \sum_{\nu\in\mathrm{A}} \sum_{\sigma\in\mathrm{B}} P_{\nu\sigma} \langle\chi_\mu\chi_\nu|\chi_\lambda\chi_\sigma\rangle \tag{2.23}$$

The two-electron integrals are approximated as before (multipole) while the 1-electron terms are approximated by the extended Hükel approach, where the $\beta$s are empirical parameters (resonance integrals). The Fock matrix is then used the same way as regular restricted Hatree-Fock theory, to obtain the orbitals that correspond to the variational energy minimum.

To determine the parameters for semi-empirical methods, it is necessary to have experimental data to fit the parameters (*ab initio* data is also sometimes used). For MNDO, the experimental values is the heat of formation, due to the amount of known experimental data for a lot of known systems. The heat of formation is defined as the amount of energy needed to shape the molecule formation for the atoms (e.g. the heat of formation of water is the energy difference of $\mathrm{H}_2 + \frac{1}{2}\mathrm{O}_2$ and $\mathrm{H}_2\mathrm{O}$). The heat of formation can be written as:

$$\Delta H_{\mathrm{f}} = E_{\mathrm{ele}} + E_{\mathrm{nuc}} - \sum_{\mathrm{A}} E_{\mathrm{ele}}^{\mathrm{A}} + \sum_{\mathrm{A}} \Delta H_f^{\mathrm{A}} \tag{2.24}$$

where $E_{\mathrm{ele}}$ is the electronic SCF energy, $E_{\mathrm{nuc}}$ is the core-core repulsion term, $E_{\mathrm{ele}}^{\mathrm{A}}$ is the electronic energy for atom A, and $\Delta H_f^{\mathrm{A}}$ is the experimental heat of formation for atom A. The nuclear repulsion term can be described as purely monopol charge-charge interaction, and so the interaction between core A and B is the same as two *s*-orbitals

$$E_{\mathrm{nuc}}^{\mathrm{AB}} = Z_{\mathrm{A}} Z_{\mathrm{B}} \langle s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}} \rangle \tag{2.25}$$

The total repulsion is then a sum of all individual interactions. Since we know the structure and the experimental heat of formation for a given molecule, we can then fit our parameters to best reproduce that heat of formation.

During the parametrization it was found that the MNDO method tended to give covalent bonds that were too short, and it was difficult to adjust the existing parameters to fix this. Therefore, additional parameters were introduced in the nuclear repulsion energy. The additional parameters makes the core-core repulsion larger, and there by the bonds, as seen

$$E_{\mathrm{nuc}} = \sum_{\mathrm{A}} \sum_{\mathrm{B>A}} Z_{\mathrm{A}} Z_{\mathrm{B}} \langle s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}} \rangle \left(1 + e^{-\alpha_A r_{\mathrm{AB}}} + e^{-\alpha_B r_{\mathrm{AB}}}\right) \tag{2.26}$$

where the atomic specific parameter $\alpha$ are adjusted to give correct bond lengths. Except for OH and NH bonds, where the nuclear repulsion was scaled with the covalent bond distance[22].

All parameters was subsequently refitted against a larger data set and more parameters was added to give the Austin Model 1 (AM1)[24] and yet again the Parameterization Model 3 (PM3)[25], which are the two methods most commonly used, whereas MNDO is rarely used[17]. The main difference between AM1 and PM3 is that AM1's parameters were fitted with chemical intuition by Dewar and PM3 by Stewart was fitted with a more statistical approach with an error function summing over observables and taking the difference between calculated and experimental values. The chemical properties used consist of heats of formation, dipole moments, ionization potentials, and molecular geometries.

The MNDO method was again re-parameterized by Stewart to give the Parameterization Model 6 (PM6)[2], this time the training set of reference data used was considerably larger than that used in parameterizing PM3 where approximately 800 discrete species were used. In the optimization of the parameters for PM6, over 9,000 separate species were used. Thiel *et al* have shown[23] that a large increase in accuracy results when *d*-orbitals are added to the main-group elements that have the potential to be hypervalent, which was added to the NDDO formalism for PM6. Also, use of other types of reference data was found to be necessary[2], such as *ab initio* calculations, which provided a convenient source of reference data. The PM6 method included even more parameters in the core-core repulsion term:

$$E_{\mathrm{nuc}} = \sum_{\mathrm{A}} \sum_{\mathrm{B>A}} Z_{\mathrm{A}} Z_{\mathrm{B}} \langle s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}} \rangle \left(1 + x_{\mathrm{AB}} e^{-\alpha_{\mathrm{AB}}(r_{\mathrm{AB}} + 3\cdot 10^{-4} r_{\mathrm{AB}}^6)}\right) \tag{2.27}$$

where the main difference is instead of having the parameter $\alpha_{\mathrm{A}}$ for each atom type, we have the parameter $\alpha_{\mathrm{AB}}$ for each atom-atom pair, which increases the amount of parameters greatly. Core-core specific terms was also included for OH, NH, CH and CC bonds.

# Chapter 3

# Correcting the Electronic Energy

The Hatree-Fock model can typically account for approximately 99% of the total energy, however the remaining energy is where most chemical reaction lies. With increasing number of basis functions the HF model should converge to what is called the Hatree-Fock limit which is always higher than the exact energy ($E_{\text{exact}}$). The difference between the exact and the HF limit is what is known as the **correlation energy**. The correlation interaction between the electrons is described by the deviations from the Hatree-Fock approximation due to the instantaneous interaction between the electrons, rather than only the average repulsion[16, 19, 17]. The goal of correlated methods (such as Møller Plesset Pertubation theort and Coupled Cluster) for solving the electronic energy is to calculate the remaining correlation due to the electron-electron interaction, which is especially important for non-covalent interactions.

However, these correlated methods are computationally costly and will not scale well to the system sizes of bio-chemical importance. Instead the trend is to use fitted empirical models to correct the energy for correlation energy effects such as hydrogen bonds and dispersion effects. Such methods as PM6-DH2[3], PM6-DH+[5], PM6-D3H4[4] and HF-3c[13]. The last being the new corrected Hartree-Fock method by Sure and Grimme, which consists of 3 empirical correction to the HF energy.

## 3.1 Correcting for Dispersion

Dispersion interaction (also called **London interaction**) is the intermolecular attraction due to correlated fluctuations in the electron density on neighbouring molecules. The dispersion interaction is present for all molecules and is dominant for non-polar molecules. It is a contribution to the **van der Waals interaction**, and those that vary with separation as 1/r$^6$. The strength of the dispersion interaction interaction is closely related to the polarizability of the molecule, which arises from coupling of instantaneous fluctuations in the charge distribution on two neighbouring molecules. These fluctuations can give rise to an instantaneous dipole, which may induce a dipole back to the neighbouring molecule and given the orientations of the two dipole can give rise to an attractive interaction between the two molecules[16, 26].

The most popular way to correct for dispersion interaction in a semi-empirical way, i.e. without running expensive *ab initio* calculations, is to use the third-generation dispersion correction (DFT-D3) by Stefan Grimme and co-workers[14]. The dispersion term consist of a two-body and a three-body term:

$$E_{\text{disp}}^{\text{D3}} = E_{\text{disp}}^{(2)} + E_{\text{disp}}^{(3)} \tag{3.1}$$

With the most dominant contribution to the interaction energy being the two-body term, which is given by

$$E_{\text{disp}}^{(2)} = \sum_{\text{A}\neq\text{B}} \sum_{n=6,8} s_n \frac{C_n^{\text{AB}}}{r_{\text{AB}}^n} f_{\text{damp},n} \tag{3.2}$$

here the first sum is over all atom pairs in the system and the second sum is over the $n$th-order dispersion coefficient $C_n^{\mathrm{AB}}$. In the third-generation these coefficients are calculated in a *ab initio* way, instead of empirically derived, as the previous versions[14]. $s_6$ and $s_8$ is a scaling factors (fitted parameters), for the specific method or functional. Usually $s_6$ is set to unity. $r_{\mathrm{AB}}$ is the internuclear distance. In order to avoid near singularities for small $r_{\mathrm{AB}}$, the damping function $f_{\mathrm{damp},n}$ is introduced which determines the range of the dispersion correction. The damping function can be either zero-damping (ZD) or Becke and Johnson (BJ) type damping.

The role of the $f_{\mathrm{damp}}$ function is to damp the dispersion contribution to zero for short ranges. The zero-damping function is chosen because Grimme *et al.* found it to be numerically stable and convenient also for higher dispersion orders.

$$f_{\mathrm{ZD},n} = \frac{1}{1 + 6(r_{\mathrm{AB}}/(s_{r,n}R_0^{\mathrm{AB}}))^{-\alpha_n}} \tag{3.3}$$

where $s_{r,n}$ is the order-dependent scaling factor of the cutoff radii $R_{\mathrm{AB}}^0$. The cutoff radii is defined as

$$R_0^{\mathrm{AB}} = \sqrt{\frac{C_8^{\mathrm{AB}}}{C_6^{\mathrm{AB}}}} \tag{3.4}$$

The other possibility is to use the rational damping as proposed by Becke and Johnson

$$f_{\mathrm{BJ},n} = \frac{1}{(a_1 R_0^{\mathrm{AB}} + a_2)^n} \tag{3.5}$$

where $R_0^{\mathrm{AB}}$ is the cutoff radii. $a_1$ and $a_2$ are fitted parameters. The three-body contribution has a small effect on medium-sized molecules. The three-body energy term $E_{\mathrm{disp}}^{(3)}$ for the atoms A, B and C is defined as

$$E_{\mathrm{disp}}^{(3)} = -\frac{1}{6} \sum_{\mathrm{A}\neq\mathrm{B}\neq\mathrm{C}} \frac{C_9^{\mathrm{ABC}}(3\cos\theta_a \cos\theta_b \cos\theta_c + 1)}{(r_{\mathrm{AB}} r_{\mathrm{BC}} r_{\mathrm{CA}})^3} f_{\mathrm{damp}}^{(3)} \tag{3.6}$$

where $\theta_a$, $\theta_b$ and $\theta_c$ are the internal angles of the triangle formed by $r_{\mathrm{AB}}$, $r_{\mathrm{BC}}$ and $r_{\mathrm{CA}}$. The $C_9$ dispersion coefficient is approximated by

$$C_9^{\mathrm{ABC}} \approx -\sqrt{C_6^{\mathrm{AB}} C_6^{\mathrm{BC}} C_6^{\mathrm{CA}}} \tag{3.7}$$

The damping function $f_{\mathrm{damp}}^{(3)}$ is similar to the zero-damping function, eq. 3.3.

## 3.2 Correcting for Hydrogen Bonds

Hydrogen bonds are the electrostatic attractive interaction between molecules in which a hydrogen is bound to a electronegative atom (donor) and is in the vicinity of another electronegative atom (accepter) with a lone pair of electrons. Just like dispersion, there exists different empirical correction schemes for correcting hydrogen bond interaction energies[5, 3, 4]. Most notably is the H+[5], H2[3] and H4[4] introduced by Korth, Hobza and co-workers. Common for all the hydrogen bonding schemes is the inclusion of penalty/reward functionality of the hydrogen bonding angles of the configuration.

We will only focus on the methodology of the third-generation hydrogen bonding correction H+. As part of this master thesis the H+ module was implemented in GAMESS. The correction energy $E(\mathrm{H+})$ is given by:

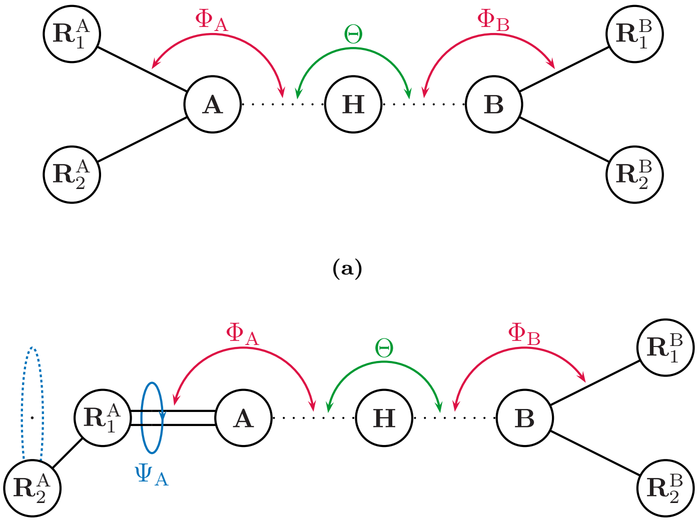
$$E(\mathrm{H+}) = \sum_{AB} \frac{C_{\mathrm{A}} + C_{\mathrm{B}}}{2r_{AB}^2} \cdot f_{\mathrm{geom}} \cdot f_{\mathrm{bond}} \cdot f_{\mathrm{damp}} \tag{3.8}$$

where the sum runs over all hydrogen bonds involving N and O atoms. $r_{\mathrm{AB}}$ is the donor-acceptor distance for the given hydrogen bond geometry, with A and B being the two possible acceptor/donor electronegative atoms, either oxygen or nitrogen. $C_{\mathrm{A}}$ and $C_{\mathrm{B}}$ are adjustable parameters and refer to either $C_{\mathrm{N}}$ and $C_{\mathrm{O}}$.

The geometrical term $f_{\mathrm{geom}}$ is defined as

$$f_{\mathrm{geom}} = \cos^2\theta \cdot \cos^2\phi_{\mathrm{A}} \cdot \cos^2\psi_{\mathrm{A}} \cdot \cos^2\phi_{\mathrm{B}} \cdot \cos^2\psi_{\mathrm{B}} \tag{3.9}$$

where $\theta$ is the angle defined by atom A, atom B and the hydrogen (see Figures 3.1a and 3.1b). The angle $\phi$, and torsion angle $\psi$ are both defined by the hydrogen bonding geometry. The angles $\phi$ are calculated from the difference between the target angle $\phi_{\mathrm{target}}$ and the present bond angle in the complex $\Phi_{\mathrm{X}}$. The target angle $\phi_{\mathrm{target}}$ is the optimum angle for hydrogen bonds. Target angles are defined in a complicated heuristic fashion, please see the source code posted on GitHub for more details[27]. The torsion angles $\psi$ are defined similarly and calculated as the difference between target dihedral angle and the structural angle $\Psi$. Where $\Psi_{\mathrm{X}}$ is the dihedral angle between $\mathrm{R_1R_2X\cdots H}$, which is used for both the donor and acceptor as seen in Figures 3.1a and 3.1b. Here $\mathrm{R}_1$ is defined as the $\mathrm{R}_x$ closest to the hydrogen.


$\Phi_{\mathrm{A}}$
$\Theta$
$\Phi_{\mathrm{B}}$
$\mathrm{R}_1^{\mathrm{A}}$
$\mathrm{R}_2^{\mathrm{A}}$
A
H
B
$\mathrm{R}_1^{\mathrm{B}}$
$\mathrm{R}_2^{\mathrm{B}}$
$\Psi_{\mathrm{A}}$

**(a)**

**(b)**

**Figure 3.1:** Illustrating the angles of the H+ model when the hydrogen bond acceptor is $\mathrm{sp}^3$ (a) and $\mathrm{sp}^2$ (b) hybridized. $\Theta$ is the angle between atoms A and B. $\Phi_X$ is the angle between the hydrogen and the $\mathrm{R}_1$ atom, $\mathrm{H{\cdot\cdot}X\text{-}R_1}$, where $\mathrm{R}_1$ is the atom closest to the H atom. $\Psi_X$ is the dihedral angle between $\mathrm{R_1R_2X\cdots H}$.

The bond damping function $f_{\mathrm{bond}}$ is defined as:

$$f_{\mathrm{bond}} = 1 - \frac{1}{1 + \exp[-60 \cdot (r_{\mathrm{XH}}/1.2 - 1)]} \tag{3.10}$$

where $r_{\mathrm{XH}}$ is the distance between the hydrogen atom and the donor atom, which is defined as the shorter one of the distances $r_{\mathrm{AH}}$ and $r_{\mathrm{BH}}$. The damping function $f_{\mathrm{damp}}$ is defined as:

$$f_{\mathrm{damp}} = \left(\frac{1}{1 + \exp[-100 \cdot (r_{\mathrm{AB}}/2.4 - 1)]}\right)\left(1 - \frac{1}{1 + \exp[-10 \cdot (r_{\mathrm{AB}}/7.0 - 1)]}\right) \tag{3.11}$$

where $r_{\mathrm{AB}}$ is the distance between the two electronegative atoms A and B.

The E(H+) implementation differs slightly from the one originally proposed by Korth[5]. Changes were made to avoid problems with optimization of hydrogen bond complexes involving particular configurations, including especially ketone (C=O) groups interacting with amide-like (NR3) groups.

In the original implementation, optimization problems can originate from target angle calculation based on the torsion angle of the NR3 group. Target angles are the optimal (text-book) angles for a given H-bond arrangement. H-bond energies are computed based on the deviation of all angular coordinates from their respective target (optimal) angles, see reference [28] for a detailed explanation. The target angle would switch during optimization steps as the definition of the torsion angle would switch, and never find a minimum, as the torsion angle is defined as seen in Figure 3.1a. The model was updated with new target angles for tetragonal NR3 configuration case, and the estimation of target angles for NR3 groups now based on the hydrogen bonding configuration (with a double bond indicating a planar structure).

The H+ model is implemented including the analytical derivative. The analytical gradient is done using internal coordinates, angles, torsions and distances, from the energy model eq. 3.8, which is then converted to Cartesian atomic coordinates by a conversion algorithm. The source code for this module has been sent to the official GAMESS version, which will be available soon and is also available as a stand-alone module on GitHub[27].

## 3.3 Correcting for Basis Set Size

Using the restricted Hartree-Fock (i.e. without NDDO approximation) with a small basis, overbinding between a dimer will occur, which is a general phenomenon known as **Basis Set Superposition Error** (BSSE). For small basis sets each monomer uses basis functions located on the other monomer in the dimer. This is due to the decrease in internal energy, which in turn leads to an overestimation of the interaction strength. **Counterpoise correction** is a method to limit the error that results when studying an intermolecular reaction using an incomplete basis set. To get the corrected energy, calculations are done for each monomer with added basis functions localized on the other monomers location, but without including the nuclei or the electron of the other monomer. That way we can correct the energy by subtracting the overbinding effect for the basis set.

An adaptation of this method has been created, by Kruse and Grimme, the geometrical counterpoise correction (gCP)[29], in a semi-empirical way. The method depends only on the molecular geometry, i.e., no input from the electronic wave-function is required and hence is applicable to large molecules. The gCP empirical correction term for counterpoise correction is defined as

$$E_{\mathrm{BSSE}}^{\mathrm{gCP}} = \sigma \sum_{A}^{N} \sum_{A \neq B}^{N} E_{\mathrm{A}}^{\mathrm{miss}} \frac{\exp\left(-\alpha (R_{\mathrm{AB}})^{\beta}\right)}{\sqrt{S_{\mathrm{AB}} N_{\mathrm{B}}^{\mathrm{virt}}}} \tag{3.12}$$

where $\alpha$, $\beta$ and $\sigma$ are fitting parameters, $S_{\mathrm{AB}}$ is a Slater-type overlap integral and $N_{\mathrm{B}}^{\mathrm{virt}}$ is the number of virtual orbitals on B in the target basis. The $S_{\mathrm{AB}}$ is evaluated over a single $s$-type orbital centered on each atom and using optimized Slater exponents. The gCP parameters were fitted in a least-square sense against counterpoise correction data[29].

To further correct the electronic energy for basis set deficiencies when using small basis sets another correction term is introduced by Sure and Grimme[13], for systematically overestimated covalent bond lengths for electronegative elements and is again calculated as a sum over all atom pairs. The correction term is defined as $E_{\mathrm{SRB}}$:

$$E_{\mathrm{SRB}} = -s \sum_{\mathrm{A}}^{N} \sum_{\mathrm{A} \neq \mathrm{B}}^{N} (Z_{\mathrm{A}} Z_{\mathrm{B}})^{3/2} \exp\left(-\gamma (R_0^{\mathrm{AB}})^{3/4} R_{\mathrm{AB}}\right) \tag{3.13}$$

Where $R_0^{\mathrm{AB}}$ is the default cutoff radii (eq. 3.4) as determined by *ab initio* for D3 dispersion correction scheme[14], $Z$ is the nuclear charge, and $s$ and $\gamma$ are fitted parameters which are set to 0.03 and 0.7 respectively for the HF-3c method. The parameters $s$ and $\gamma$ were fitted to produce the missing atomic forces of the HF-3c method for the B3LYP-D3(BJ)/def2-TZVPP equilibrium structures of 107 small organic molecules[13]. The other correction ($E_{\mathrm{D3}}$ and $E_{\mathrm{gCP}}$) terms were included in the fitting procedure of $E_{\mathrm{SRB}}$, which was carried out by minimizing the HF-3c RMS gradient for the reference geometries.

# Chapter 4

# Checking the Electronic Energy

We have introduced the methodology of two approaches, namely the standard restricted Hartree-Fock and the approximately Hartree-Fock approach, NDDO (PM6). As explained earlier, these methods are lacking in calculating non-covalent interactions, which was why we introduced several correction schemes. Now that we have corrected the original energy, we want to check if it gives the correct energy. Usually we do not care about the absolute electronic energy, but rather relative energies, such as energy barriers or interaction energies. To check if the energy model we use is accurate enough to explain interactions other than the average electron-electron repulsion, we compare interaction energies calculated with a high-level correlated method, such as CCSD(T)/CBS. A very nice database has been created for just this reason, which contains geometries optimized at the MP2/cc-pVTZ level and single point calculation at the CCSD(T)/CBS level[30]. Different sets exists, but primarily semi-empirical models use the S22 and the S66 complex set, where we can find different complexes with focus on bio-chemical interaction, e.g. dispersion and hydrogen-bonding, e.g. Figure 4.1 where both dispersion and hydrogen-bonding takes place.

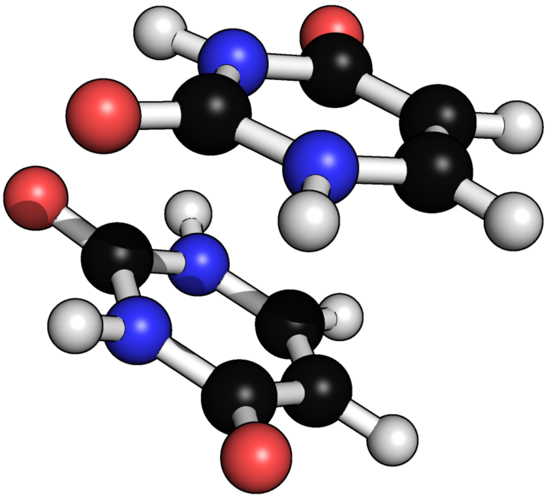


**Figure 4.1:** Complex no. 13 of the S22 set, an Uracil dimer in stack formation, showing both dispersion and hydrogen bond interaction.

## 4.1 Computational Details

During this study we primarily work with two different quantum packages, namely GAMESS[12] and MOPAC[31]. It is important to know which one are used for the different methods, as the methods are very much coupled with the way it is implemented and the associated algorithms for geometrical optimizations. MOPAC is close-sourced which is what prompted the motivation

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.799v1 | CC-BY 4.0 Open Access | rec: 16 Jan 2015, publ: 16 Jan 2015

to re-implement PM6 in GAMESS, as GAMESS has several ways to control optimizations that MOPAC does not have.

As part of this study we present new dispersion and hydrogen bond corrections to the PM6 method, PM6-D3H+[11], and its implementation in the GAMESS program. The method combines the DFT-D3[14] dispersion correction (eq. 3.1) with zero-damping (eq. 3.3) with a modified version of the H+ hydrogen bond correction[5] (eq. 3.8).

$$E_{\mathrm{PM6-D3H+}} = E_{\mathrm{PM6}} + E_{\mathrm{D3}} + E_{\mathrm{H+}} \tag{4.1}$$

All NDDO based methods (including the newly implemented PM6) in GAMESS is currently only implemented with numerical gradients. The gradient of the dispersion correction is evaluated numerically, by using a centered finite difference scheme, for three-body calculations, and analytically for two-body calculations. The analytical three-body gradient is published but not yet implemented in GAMESS. For the hydrogen bond term the analytical gradient is used. For Hessian calculations in GAMESS we use double displacement (`NVIB=2` in the `$force` group in the GAMESS input file).

The HF-3c method[13] is the recent developed semi-empirical corrected Hartree-Fock method, introduced by Sure and Grimme, using a very small basis set (MINIX), and using three of the correction terms introduced earlier, namely dispersion (eq. 3.1), BSSE (eq. 3.12) and a short range term (eq. 3.13).

$$E_{\mathrm{HF-3c}} = E_{\mathrm{HF}}^{\mathrm{MINIX}} + E_{\mathrm{D3}} + E_{\mathrm{gCP}} + E_{\mathrm{SRB}} \tag{4.2}$$

The Hartree-Fock method is, as mentioned earlier, much slower than the NDDO based PM6, however as part of this work the HF-3c method has been implemented in GAMESS and coupled with the Fragment Molecular Orbital (FMO)[32] scheme, FMO-HF-3c (unpublished). This makes the method able to scale well to large system sizes. For the FMO-HF-3c calculations we used only the two-body scheme of FMO.

All PM6-D3H+, HF-3c, and FMO-HF-3c calculations were done with a locally modified version of GAMESS. The source-code for the method PM6-D3H+ has been formatted and sent to the official GAMESS group in Iowa, and will be available in the official version later this year. The FMO-HF-3c interface will soon be pushed to GAMESS as well.

The semi-empirical methods PM6[2], PM6-DH2[3] and PM6-DH+[5] are used as implemented in the closed-sourced program MOPAC. All MOPAC calculations were done with MOPAC2012[31, 33] version of MOPAC. Geometry optimizations were done with the LBFGS optimizer for reasons described later, unless noted otherwise. The COSMO model[34] were used to model bulk solvation for the protein calculations.

To benchmark and test our implementations, we performed various calculations on the S22[35] and S66[36] set of complexes from the Benchmark Energy and Geometry Database (BEGDB)[30]. The BEGDB database contains structures and corresponding interaction energies calculated at the MP2/cc-pVTZ and estimated CCSD(T)/CBS level of theory, respectively.

Geometry optimizations of the complexes in S22 and S66 were done with a variety of convergence criteria which will be discussed in detail later. Geometry optimizations of Chignolin (PDB: 1UAO) and the Tryptophan-cage (PDB: 1L2Y) using PM6-DH+, PM6-D3H+ and FMO-HF-3c were also carried out. We used the first structure available in each of the downloaded structure files. For comparison, we performed two-body Fragment Molecular Orbital (FMO)[32] geometry optimizations using RHF/6-31G(d)[37, 38, 39, 40] and the D3 dispersion correction[14, 41].

Calculations were performed in the gas phase and in bulk solvent using a polarizable continuum to model the solvent.[42] For solvated PM6-D3H+ calculations, we used a recent C-PCM implementation[43] for SQM methods in GAMESS. For the FMO calculations, we used the recent completely analytical RHF/C-PCM gradient[44]. All PCM calculations were done using the

FIXPVA[45] tesselation scheme with 60 tesserae per sphere. All geometry optimizations used a convergence criterion of $5.0 \times 10^{-4}$ Hartree/Bohr, unless noted otherwise.

Timings was carried out on either a 8 core Intel(R) Xeon(R) CPU X5560 @ 2.80GHz or 24 core AMD Opteron(tm) Processor 6172 @ 2.1 GHz machine.

## 4.2 Parameterization

For the new method in GAMESS, PM6-D3H+, new parameters was needed for the two correction terms, dispersion and hydrogen-bonding. $E_{\mathrm{D3}}$ is the third generation dispersion correction developed by Grimme *et al.*, DFT-D3[14] and implemented in GAMESS by R. Peverati. Unless otherwise noted $E_{\mathrm{D3}}$ refers to the pair-wise additive dispersion correction as proposed in reference [14]. Only the zero-damping version was used, with dispersion order 6 and 8. The fitting parameters are those obtained by Grimme for PM6[9]. As described by Grimme, the parameter $s_6$ is set to unity, $\alpha$ was set to its default value. $s_8$ and the scaling parameter $s_{r,6}$ of the atomic cut-off radii used in the dispersion damping function are fitted parameters as in standard DFT-D3 (see Table 4.1 for parameters). Thus only $s_8$ and $s_{r,6}$ are optimized by Grimme for PM6-D3H, which is also used for PM6-D3H+.

Because we use a different dispersion energy function than in the previous DH+ model and make modification to the original hydrogen bonding correction model, it is necessary to determine new optimum values for the $C_{\mathrm{N}}$ and $C_{\mathrm{O}}$ parameters. The parameters for H+ are parameterized to minimize the root-mean-square deviation (RMSD) between the interaction energies for PM6 with dispersion correction only (PM6-D3) for a subset of complexes from the S22 and S66 data sets (1-7 and 1-23, respectively), plus the H+ term and the estimated CCSD(T)/CBS reference interaction energy. The $C_{\mathrm{N}}$ and $C_{\mathrm{O}}$ parameters are then scanned in ranges from -0.2 to 0.0, around the original optimum. A global optimum was found at $C_{\mathrm{N}}$ = -0.11 and $C_{\mathrm{O}}$ = -0.12, with a RMSD of 1.11 kcal/mol, as seen in Figure 4.2 and Table 4.2.

This was done using both two and three-body dispersion, but including three-body dispersion did not make any substantial difference in the resulting optimum, and the default was set to two-body for PM6-D3H+, because of the extra computational time associated with three-body gradient calculations. The computational cost becomes a time consuming issue for optimizing protein-sized molecules. The final set of parameters for both dispersion and hydrogen bond correction terms can be seen in Table 4.1.

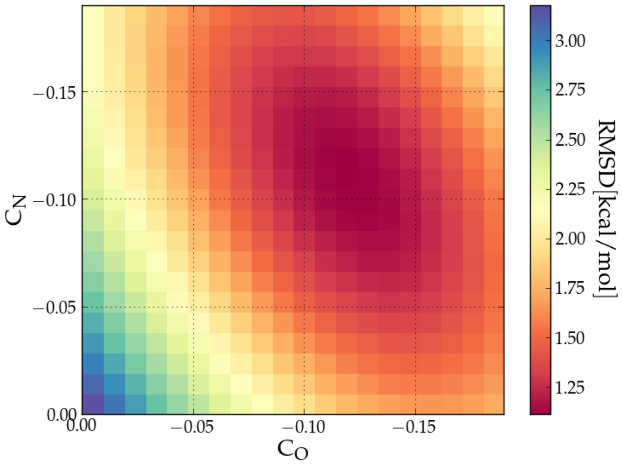


**Figure 4.2:** Scan of the two parameters for the H+ correction term, nitrogen ($C_{\mathrm{N}}$) and oxygen ($C_{\mathrm{O}}$) in the hydrogen bond dominant complexes of the S22 and S66 noncovalent complexes. A optimum was found at $C_{\mathrm{N}} = -0.11$ and $C_{\mathrm{O}} = -0.12$.

**Table 4.1:** The final parameters for the dispersion and hydrogen bond correction terms of PM6-D3H+.

| | H+ |
|---|---|
| $C_{\mathrm{N}}$ | -0.110 |
| $C_{\mathrm{O}}$ | -0.120 |
| | D3 |
| $\alpha$ | 14.000 |
| $s_6$ | 1.000 |
| $s_{r,6}$ | 1.560 |
| $s_8$ | 1.009 |

## 4.3 Interaction Energies

Table 4.2 shows results of PM6, PM6-DH+, PM6-D3H+ and HF-3c for the full, dispersion and hydrogen bond dominant complexes sets of the S22 and S66 from BEGDB. Root-mean-square deviation (RMSD), mean absolute deviation (MAD) and maximum error span (Max) with respect to the benchmark estimated CCSD(T)/CBS interaction energies are given in kcal/mol. The PM6-D3H+ method was tested using both two and three-body dispersion.

Overall, the accuracy of PM6-D3H+ is very similar to PM6-DH2 and PM6-DH+, with RMSD and MAD values within 0.02 kcal/mol of one another. The main difference is that the maximum error for PM6-D3H+ is 1.42 and 0.36 kcal/mol smaller than for PM6-DH2 and PM6-DH+, respectively. The accuracy of HF-3c is consistently better, but in the same order of magnitude as the NDDO based methods. The maximum error for PM6-DH2, -DH+, and -D3H+ were observed for the S66-19, S66-60, and S66-65 dimer, respectively and, in general, we did not notice any particular dimer that resulted in unusually large errors for all four corrections. All interaction energies can be found in supplementary information. The differences in RMSD and MAD between PM6 corrected methods are slightly larger (up to 0.13 kcal/mol) for subsets where dispersion and hydrogen-bonding dominate. The HF-3c method seems to be more consistent with reference energies, especially for hydrogen bond dominant complexes, with a RMSD of 0.5 kcal/mol smaller than -DH2, -DH+ and -D3H+, as well as a smaller max error. Including three-body dispersion correction for the PM6-D3H+ method had no substantial effect on accuracy, but might play a role for large systems.

Next, we test the methods on two sets of molecules not in the training set. Table 4.3 lists computed interaction energies for formamide dimer, pentamer-monomer, and trimer-trimer (Figure 4.3) computed with various methods. Compared to MP2/TZVP PM6-DH2 performs best for this particular system, while PM6-DH+ and PM6-D3H+ appear to perform roughly similarly, with mean absolute deviations (MAD) of 0.8 and 1.3 kcal/mol, respectively. However, it is interesting to note that the decrease in interaction energy on going from the dimer to the pentamer-monomer predicted by PM6-DH+ (3.6 kcal/mol) is somewhat lower than that predicted by other methods corrections and MP2/TZV (4.1 - 4.6 kcal/mol). This decrease comes primarily from cooperative polarization effects that are accounted for by the underlying PM6 method, and PM6, PM6-DH2, and PM6-D3H+ all predict similar decreases. It is not clear why the DH+ terms leads to an underestimation of the cooperative effect. Similarly, the HF-3c methods underestimates the binding consistently when going from dimer to trimer.

Table 4.4 contains RMSD, MAD, mean-deviation (MD) and maximum deviation relative to estimated CCSD(T)/CBS// MP2/cc-pVTZ interaction energies computed for 12 hydrogen bonded base pair complexes (List in supplementary information) from the JSCH-2005[48] set from BEGDB. The 12 complexes represent all the complexes in the JSCH-2005 set with hydrogen bonds involving N and O atoms and for which interaction energies have been computed at a level similar to that used in the parameterization of PM6-D3H+ [i.e. CCSD(T)/CBS// MP2/pVTZ]. For this set all three PM6 corrected models offer very significant increases in accuracy (e.g. a ca 8 kcal/mol decrease in the

**Table 4.2:** Root-mean-square deviation (RMSD), mean absolute deviation (MAD), as well as the maximum error (Max) with respect to the estimated CCSD(T)/CBS interaction energies from the S22 and S66 sets are presented. Hydrogen bond and dispersion subsets are complexes from S22 and S66 with a dominant factor of the interaction energy being hydrogen bond or dispersion interaction. All values are in kcal/mol.

| | ${}^{a,b}$PM6 | ${}^{b}$DH2 | ${}^{b}$DH+ | ${}^{a,c}$D3H+ | ${}^{a,d}$D3H+ | ${}^{a}$HF-3c |
|---|---|---|---|---|---|---|
| | Full set | | | | | |
| RMSD | 3.34 | 0.83 | 0.80 | 0.82 | 0.83 | 0.53 |
| MAD | 2.85 | 0.58 | 0.61 | 0.60 | 0.61 | 0.39 |
| MAX | 7.99 | 3.53 | 2.47 | 2.11 | 2.09 | 1.80 |
| | Dispersion subset | | | | | |
| RMSD | 3.15 | 0.49 | 0.49 | 0.48 | 0.54 | 0.63 |
| MAD | 2.79 | 0.42 | 0.42 | 0.36 | 0.39 | 0.48 |
| MAX | 7.29 | 0.92 | 0.92 | 1.11 | 1.43 | 1.80 |
| | Hydrogen bond subset | | | | | |
| RMSD | 4.29 | 1.05 | 0.98 | 1.11 | 1.11 | 0.58 |
| MAD | 3.65 | 0.70 | 0.80 | 0.92 | 0.91 | 0.47 |
| MAX | 7.99 | 3.53 | 2.10 | 1.85 | 1.84 | 1.36 |

[a] The calculations have been done using the GAMESS software.
[b] The calculations have been done using the MOPAC software.
[c] The calculation has been done using two-body dispersion.
[d] The calculation has been done using three-body dispersion.

**Table 4.3:** Hydrogen bond interaction energies, with various methods, from formamide dimer, pentamer-monomer, and trimer-trimer, as well as MP2/TZVP reference data. All values are in kcal/mol.

| | PM6${}^{a,b}$ | DH2${}^{b}$ | DH+${}^{b}$ | D3H+${}^{a}$ | HF-3c${}^{a}$ | MP2/TZVP${}^{d}$ |
|---|---|---|---|---|---|---|
| dimer | -5.36 | -6.71 | -7.81 | -8.12 | -6.03 | -6.65 |
| pentamer-monomer | -7.17 | -8.82 | -9.56 | -10.06 | -7.74 | -8.66 |
| trimer-trimer | -9.27 | -11.33 | -11.45 | -12.23 | -9.67 | -11.26 |

[a] The calculations have been done using the GAMESS software.
[b] The calculations have been done using the MOPAC software.
[d] From ref [46, 47].

MAD) compared to PM6. The HF-3c model also offers very similar order of accuracy in this model set compared to the other correction methods. As for the training set (Table 4.2) the accuracy of PM6-DH2, PM6-DH+, PM6-D3H+ and HF-3c are very similar, with MADs between 0.7 and 1.1 kcal/mol.

## 4.4 Molecule Geometry Optimization

All structures from the S22 and S66 data sets were optimized with PM6, and PM6-DH+ using MOPAC or PM6, PM6-D3H+ and HF-3c using GAMESS to test how well the methods reproduce the reference MP2/cc-pVTZ geometries and to compare the optimization algorithms in GAMESS and MOPAC.

For the GAMESS optimizations we used the default (quasi Newton-Raphson) geometry optimizer and defined convergence as having a maximum gradient component less than $5 \times 10^{-4}$ Hartree/Bohr and an RMS gradient less than $5/3 \times 10^{-4}$ Hartree/Bohr. These convergence criteria are five times higher than the default and are chosen because we have found that for large systems these criteria can lead to significantly faster convergence without affecting the structure or final energy significantly. For complex 58 in the S66 set it was necessary to re-compute the Hessian every

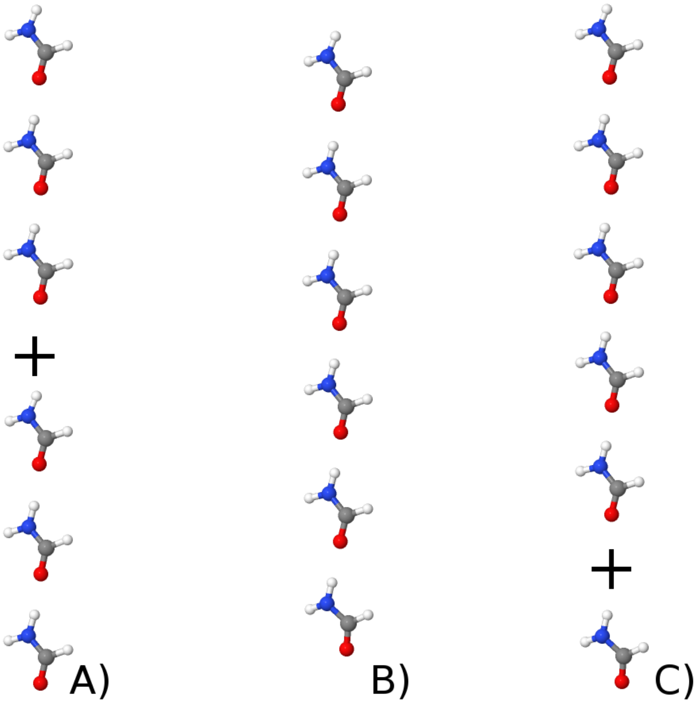


**Figure 4.3:** Illustrating the formamide trimer-trimer (a), hexamer (b) and pentamer-monomer (c).

20 steps to obtain convergence and in the case of complex 22, 51, and 58 it was necessary to skip the projection of translational and rotational degrees of freedom from the gradient to obtain convergence, which was done by settings the keyword `PROJCT=.F.` in the `$Force` group. For 11 of the complexes (see supplementary information) it was necessary to decrease convergence criterion to $10^{-4}$ Hartree/Bohr in order to remove imaginary frequencies. In the case of complex 4 and 5 from S22 PM6-D3H+ predicted that the minimum has $C_1$ symmetry rather than $C_s$ as predicted by MP2, and a deviation in the planarity structure of 0.1 Å was needed (added to the first atom). This is not the case for PM6 and thus a result of the D3H+ energy correction.

For HF-3c three structures failed to optimize because of Hessian corruption in the first optimization step. This happened for 22, 51 and 58 of the S66 set.

For the MOPAC optimization we used the LBFGS geometry optimizer because we found that this

**Table 4.4:** Root-mean-square deviation (RMSD), mean absolute deviation (MAD), mean deviation (MD), as well as the maximum error (Max) with respect to the estimated CCSD(T)/CBS interaction energies from selected complexes from JSCH-2005 dataset.

| Method | RMSD | MAD | MD | Max |
|---|---|---|---|---|
| PM6[a,b] | 8.24 | 7.98 | 7.98 | 10.71 |
| PM6-DH2[b] | 1.45 | 1.09 | 0.21 | 3.97 |
| PM6-DH+[b] | 0.94 | 0.69 | 0.46 | 1.90 |
| PM6-D3H+[a] | 1.18 | 0.95 | 0.37 | 2.45 |
| HF-3c[a] | 1.26 | 1.11 | 0.25 | 2.03 |

[a] The calculations have been done using the GAMESS software.

[b] The calculations have been done using the MOPAC software.

is the only optimization algorithm that can be practically applied to optimization of large systems. Using eigenvector following leads to termination of the geometry optimization and the following error message: ”trust radius now less than 0.00010 optimization terminating”. Based on the output the convergence criterion for the LFBGS optimizer appears to be a change heat of formation of less than ca. 0.1 kcal/mol during several consecutive optimization steps. For PM6, this convergence test was not passed after ca 200 geometry optimization steps for complex 10 and 17 from the S22 set and 29, 53, and 54 from the S66 set. For PM6-DH+, this convergence failed after ca 140 geometry optimization steps for complex 11 from the S22 set and 53, 54 and 60 from the S66 set. In all these cases MOPAC terminates the geometry optimization after the mentioned number of steps with the message: ”a failure has occurred”

The results are summarized in Table 4.5. The average RMSD between the MP2/cc-pVTZ and semi-empirical structures are below 0.28 Å for all methods and a factor of two lower for the GAMESS optimizations. The RMSD was calculated using the Kabsch algorithm[49, 50], for all the atoms, including hydrogens. For the hydrogen bonding subset RMSD was calculated for the hydrogen bond lengths, which are much lower with GAMESS, and with PM6-D3H+ being the lowest with a RMSD of 0.08 Å. The GAMESS optimizations converge, on average, in 30 steps, while the MOPAC optimization takes 10 times more steps.

**Table 4.5:** Geometry optimization of equilibrium conformations of the S22 and S66 datasets in gas phase. Root-mean-square-deviation was calculated between the optimized structures and the original structure from S22 and S66, as well as the hydrogen bond lengths. The average number of steps ($\bar{N}_S$), average of the final root-mean-squared gradient (RMS) in Hartree/Bohr, and average number of imaginary frequencies ($\bar{N}_i$) was noted for the different methods.

| | avg. RMSD [Å] | HB RMSD [Å] | $\bar{N}_S$ | avg. Gradient RMS | $\bar{N}_i$ (max) |
|---|---|---|---|---|---|
| PM6[a] | 0.11 | 0.13 | 30 | $1.0\times10^{-4}$ | 0.02 (1) |
| PM6-D3H+[a] | 0.12 | 0.08 | 31 | $1.0\times10^{-4}$ | 0.07 (1) |
| PM6[b,c] | 0.28 | 0.24 | 229 | $1.4\times10^{-3}$ | 0.71 (6) |
| PM6-DH+[b,d] | 0.21 | 0.24 | 376 | $2.3\times10^{-3}$ | 0.79 (9) |
| HF-3c[a,e] | 0.10 | 0.05 | 32 | $1.1\times10^{-4}$ | 0.51 (3) |

[a] The calculations have been done using the GAMESS software.
[b] The calculations have been done using the MOPAC software.
[c] Averages computed without complexes 10 and 17 from S22 and 29, 53 and 54 from S66, as they did not converge.
[d] Averages computed without complexes 11 from S22 and 53, 54, 60 and 63 from S66, as they did not converge.
[e] Averages computed without complexes 22, 51, and 58 from S66, as they did not converge.

Furthermore, MOPAC optimized geometries tend to have a significantly larger RMS gradient, compared to GAMESS. This leads to significantly more imaginary frequencies in a subsequent vibrational analyses compared to those obtained with GAMESS. In the case of MOPAC/PM6-DH+ 54, 17, 15, and 2 geometries result in 0, 1, 2, and ≥ 3 imaginary frequencies, while the corresponding numbers for GAMESS/PM6-D3H+ are 82, 6, 0, and 0 and 56, 23, 6, 3 for GAMESS/HF-3c. Re-optimizations with a higher convergence criteria for HF-3c has not yet been done, but optimizing the complexes with higher criteria, as with PM6-D3H+, could remove a lot of the imaginary frequencies. Using the (default) eigenvector following algorithm in MOPAC for comparison results in 60, 19, 5, and 4 geometries with 0, 1, 2, and ≥ 3 imaginary frequencies, respectively, with complexes 1 and 3 from S22 and 1 and 20 from S66 failing the optimization.

For four of the six cases where a PM6-based GAMESS optimization leads to a structure with a single imaginary frequency a convergence criterion of $10^{-4}$ Hartree/Bohr is used, but lowering the convergence criterion further does not remove the imaginary frequencies. In the sixth case, complex 16 in the S66 set (water hydrogen bonded to an amide group - Figure 4.4), the optimization stalls, when setting convergence criterion to $10^{-4}$ Hartree/Bohr, with the maximum gradient oscillating between $3 \times 10^{-4}$ and $2 \times 10^{-4}$ Hartree/Bohr. This is due to the dihedral angle $\psi$ (Eq. 3.9) which is defined as $R_1 R_2 X \cdots H$ (cf. Figures 3.1a and 3.1b), where $R_1$ is defined as the atom closest to the H atom. In the case of the amide-water hydrogen bond, $R_1$ and $R_2$ are the two water H atoms, which are approximately equidistant from the amide proton. The oscillation in the maximum gradient is caused by the oscillation between two different definitions of $\psi$, which has an effect on the gradient direction. The normal mode associated with the imaginary frequency for the structure converged with a convergence criterion of $5 \times 10^{-4}$ corresponds to a motion between these two structures, so this is likely the explanation for the imaginary frequency. Similarly, in the case of the complex 1 in the S22 set (ammonia dimer), we believe the imaginary frequency is due to highly symmetric hydrogen configuration, with switching torsion angles (atomic definition of $\psi$). Since this only affects structures with highly symmetric hydrogen bonds it is unlikely to cause problems in most applications. We note that the PM6-DH+ method has the same problem.

In the remaining four cases where a GAMESS optimization leads to a structure with an imaginary frequency the cause is most likely an extremely flat potential energy surface for the corresponding degrees of freedom: all imaginary frequencies are $< 31i$ cm$^{-1}$. Similarly, the lowest real frequencies for these five cases are all $< 40$ cm$^{-1}$.

In summary, the PM6-D3H+ method as implemented in GAMESS offers an attractive alternative to PM6-DH+ in MOPAC in cases where the default geometry optimizer fails to find a converged structure and the LBFGS optimizer must be used and a vibrational analysis is needed e.g. when computing vibrational free energies.

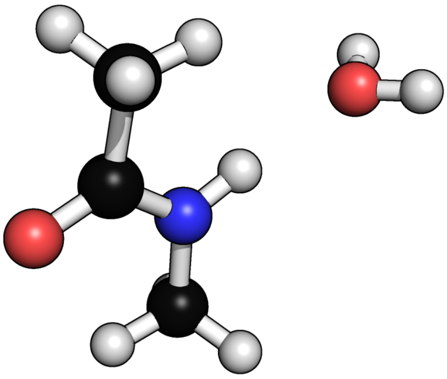


**Figure 4.4:** Hydrogen bond configuration of complex 16 of the S66 set.

## 4.5 Fragmenting the Energy

The HF-3c method was interfaced to the fragment molecular orbitals (FMO) method in GAMESS. For all FMO calculations the RCORSD and RESDIM keywords are set to 1.75 Å(default is 2.0), which is the cutoff for approximating the SCF energy by electrostatic interaction. However, as seen in table 4.6, this also affects the correction terms in HF-3c, and not only the SCF energy. This is because the implementation of 3c in GAMESS is based on already implemented interface between the DFTDx method and FMO, as the two extra corrections are called through the DFDx subroutine (for easy FMO interfacing). The semi-empirical correction terms should not be cut of for any cutoff distance, as the electrostatic interaction is an approximation to the HF energy, and does not include the energy of dispersion interaction of the correction term. Disabling the RCORSD and RESDIM keywords (setting them to zero) resets the correction energy to be exactly the same as with a standard HF-3c calculation. All FMO calculation was done using RCORSD and RESDIM.

**Table 4.6:** The total, SCF and correction energy of the HF-3c method and FMO2-HF-3c interface, as printed in the GAMESS log file on a cluster of 5 water molecules. Energies are in Hartree.

| Method | $E_{\text{total}}$ | $E_{\text{SCF}}$ | $E_{\text{3c}}$ |
|---|---:|---:|---:|
| FMO2-HF-3c | -377.559431948 | -377.525994904 | -0.033437044 |
| FMO2-HF-3c[a] | -377.559531754 | -377.525995106 | -0.033536648 |
| HF-3c | -377.559523999 | -377.525987351 | -0.033536648 |

[a] Calculations was done with RCORSD and RESDIM disabled.

## 4.6 Protein Structure Refinement

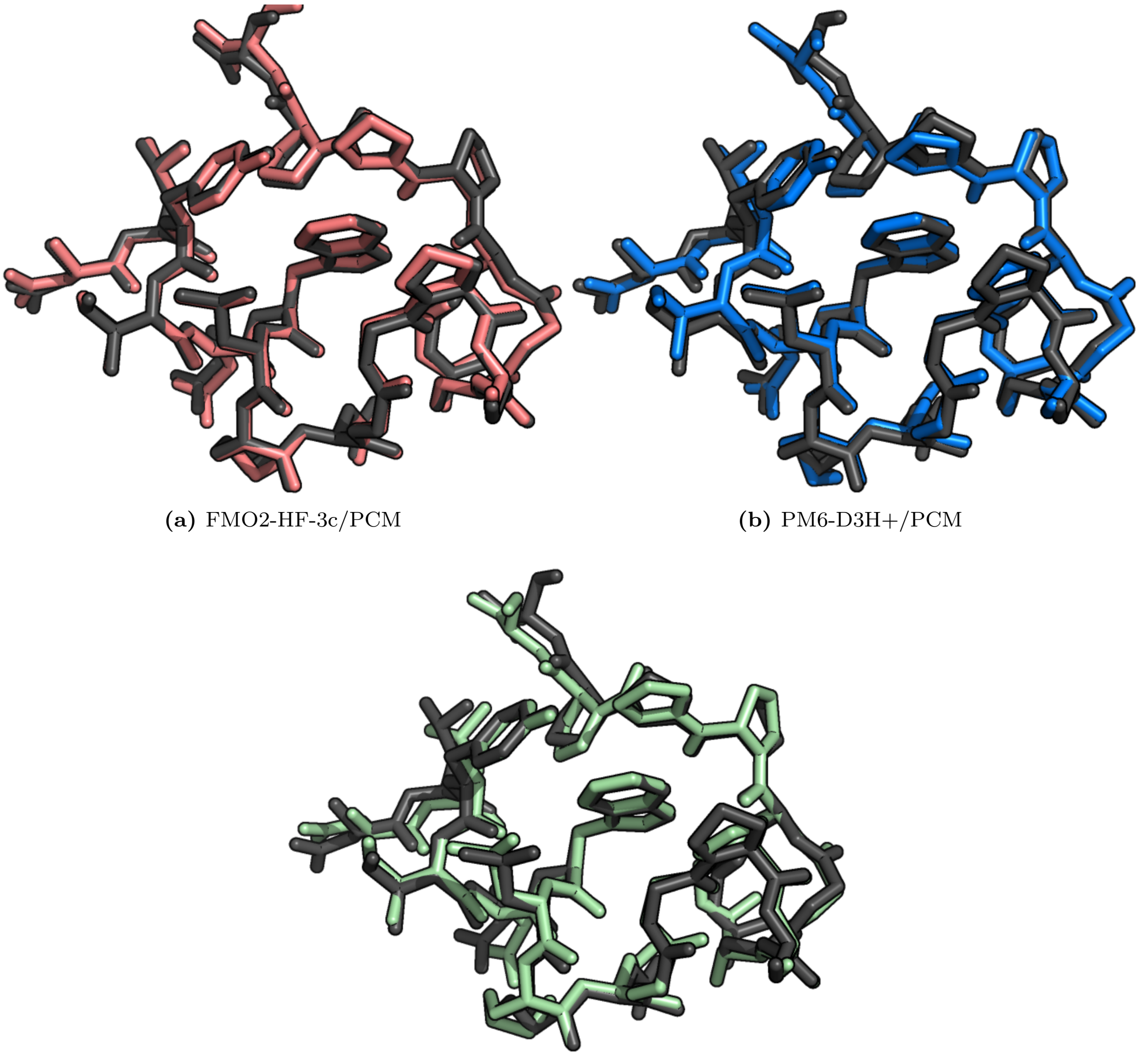
In this section we test the applicability of the PM6-D3H+ and the FMO2-HF-3c methods, combined with the PCM for bulk solvation as implemented in GAMESS, to geometry optimization of large systems such as proteins and compare to corresponding calculations performed using MOPAC.

We optimize the proteins Chignolin (1UAO) and Trp-Cage (1L2Y), which are two small proteins with 138 and 304 atoms, respectively. The optimized semi-empirical structures are compared to the reference structure optimized at the RHF/6-31G(d) level of theory using dispersion correction (DFTD3) and two-body Fragment Molecular Method (FMO2). Previous calculations by Nagata *et al.*[44] have shown that this level of theory yields protein structures in good agreement with corresponding MP2 calculations. Optimized reference structures are available on GitHub[51].

The results are summarized in Table 4.7. The RMSD values are about 1 Å in the gas phase for both PM6 methods, with PM6-DH+ being slightly smaller. The RMSD values for the structures in solution are slightly larger compared to the corresponding gas phase values for PM6-DH+, and slightly smaller for PM6-D3H+. For Trp-cage both PM6 methods converge in about half the number of steps in solvent compared to gas phase. The structural overlap between FMO2-HF-3c/PCM, PM6-D3H+/PCM and PM6-DH+/COSMO optimizations and the reference structure can been seen in Figures 4.5 and 4.6.

MOPAC requires significantly more optimization steps than GAMESS to converge, but the overall time for optimization of the structures is by far faster than GAMESS. The difference in CPU time per geometry optimization step is significantly larger for optimization in bulk solvent, which indicates that it is the difference in the COSMO and PCM interfaces that differ most in terms of CPU requirements. Despite being significantly slower than PM6-DH+/COSMO, the PM6-D3H+/PCM implementation in GAMESS is sufficiently fast to make geometry optimizations of small proteins feasible. The FMO2-HF-3c method is more computational demanding than the PM6 based methods and even though the Trp-Cage structure converges in 100 less steps than for the Chignolin, the time it takes is still significantly more. However the calculations for HF-3c was only done on 8 cores, and the FMO methods makes it possible to have great scaling over large numbers of CPUs. However, spite being the slowest, the FMO2-HF-3c finds the structures closets to reference structure with a RMSD of 0.5$\AA$ lower than PM6-D3H+ in solvent.

The number of imaginary frequencies computed for the optimized protein geometries ($N_i$) are listed in Table 6, using the PM6 methods. Hessian calculations was not done for the FMO2-HF-3c method. Again, the GAMESS optimization leads to significantly fewer imaginary frequencies: 3 and 2 using PM6-D3H+/PCM implemented in GAMESS, compared to 5 and 12 for Chignolin and Trp-cage using PM6-DH+/COSMO implemented in MOPAC. In the case of GAMESS the number of imaginary frequencies can be reduced to 0 for both proteins by decreasing the geometry optimization criterion (OPTTOL) to $1 \times 10^{-4}$ aus. This required 205 and 298 additional optimization steps for Chignolin and Trp-cage, respectively.

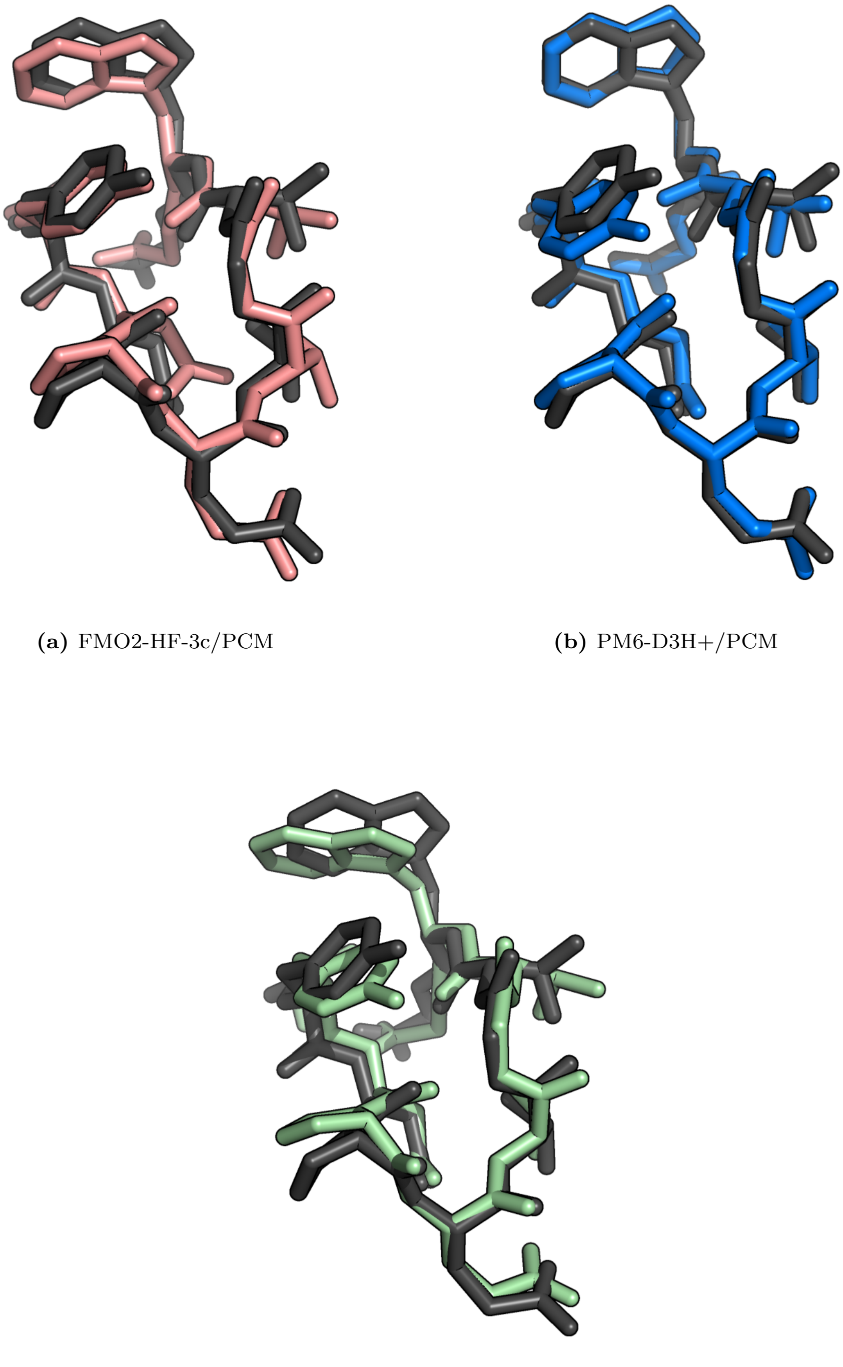


**(a)** FMO2-HF-3c/PCM

**(b)** PM6-D3H+/PCM

**(c)** PM6-DH+

**Figure 4.5:** Trp-cage (1L2Y) optimized with FMO2-RHF-D3/6-31G(d)/PCM (black), compared to (a) FMO2-HF-3c/PCM (red), (b) PM6-D3H+/PCM (blue) and (c) PM6-DH+/COSMO (green). This figure was made with PyMol[52].



(a) FMO2-HF-3c/PCM

(b) PM6-D3H+/PCM

(c) PM6-DH+

**Figure 4.6:** Chignolin (1UOA) optimized with FMO2-RHF-D3/6-31G(d)/PCM (black), compared to (a) FMO2-HF-3c/PCM (red), (b) PM6-D3H+/PCM (blue) and (c) PM6-DH+/COSMO (green). This figure was made with PyMol[52].

**Table 4.7:** Optimized proteins Chignolin with 138 atoms and Trp-Cage with 304 atoms, in gasphase and implicit solvent, using PM6-DH+, PM6-D3H+ and HF-3c with COSMO, PCM and PCM respectively for solvent polarization. RMSD (in Å) are calculated with reference to the protein structures optimized at FMO2-RHF-D3/6-31G(d) level of theory and FMO2-RHF-D3/6-31G(d)/PCM level for solvent effects. Time in hours and number of optimization steps were noted. Calculations was run on a single core, except for HF-3c which was run on 8 cores.

| | | Solvent | | | | Gasphase | | | |
|---|---|---|---|---|---|---|---|---|---|
| System | PDB | RMSD [Å] | Time [h] | Steps | $N_i$[a] | RMSD [Å] | Time [h] | Steps | $N_i$[a] |
| | | | | PM6-DH+ | | | | | |
| Chignolin | 1UAO | 1.14 | 0.1 | 941 | 5 | 0.90 | 0.1 | 739 | 4 |
| Trp-Cage | 1L2Y | 1.23 | 0.6 | 882 | 12 | 1.89 | 1.1 | 1774 | 2 |
| | | | | PM6-D3H+ | | | | | |
| Chignolin | 1UAO | 0.56 | 0.6 | 128 | 3 (0) | 0.98 | 0.2 | 204 | 0 (0) |
| Trp-Cage | 1L2Y | 0.83 | 5.2 | 174 | 2 (0) | 1.61 | 5.4 | 481 | 2 (0) |
| | | | | FMO2-HF-3c | | | | | |
| Chignolin | 1UAO | 0.83 | 27.7[b] | 186 | n/a | 1.07 | 28.1[b] | 262 | n/a |
| Trp-Cage | 1L2Y | 0.35 | 44.1[b] | 88 | n/a | 1.09 | 104.5[b] | 248 | n/a |

[a] Number of imaginary frequencies for OPTTOL = $5 \times 10^{-4}$ ($1 \times 10^{-4}$) aus.

[b] Calculations was done using 8 cores.

The relative speedup from running in parallel in solvent is shown on Figure 4.7, where no improvement is observed beyond 8 cores for all PM6-based methods. The timings were done on 24 core AMD Opteron(tm) Processor 6172 @ 2.1 GHz machine for GAMESS and 8 core Intel(R) Xeon(R) CPU X5560 @ 2.80GHz for MOPAC, because we were unable to get MOPAC running on the AMD ones. Using the dispersion correction and hydrogen bond correction on the PM6 method in GAMESS reduces the relative speedup from 4 to about 2. The correction terms to the PM6 energy only runs in serial, and a modest speedup could be gained by parallelizing them. Here we note that the poor scaling of run times with regards to the number of CPUs used is an inherent problem for semi-empirical since the matrix diagonalization in the SCF procedure cannot be efficiently parallelized[33].

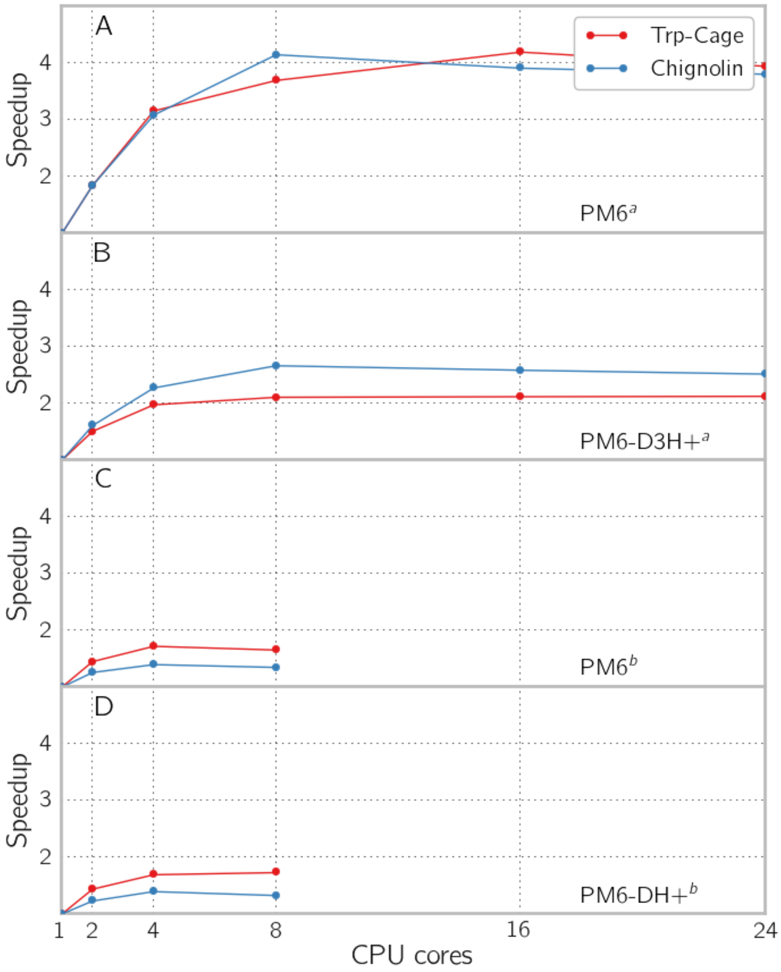


**Figure 4.7:** Speedup by using multiple cores with solvent enabled for single point energy and gradient evaluation of the proteins Trp-Cage (1L2Y) with 304 atoms and Chignolin (1UAO) with 138 atoms, using (A) PM6 and (B) PM6-D3H+ in GAMESS and (C) PM6 and (D) PM6-DH+ in MOPAC. The evaluation was done using implicit solvent models COSMO and PCM for respectively MOPAC and GAMESS. $^a$ The calculations have been done using the GAMESS software. $^b$ The calculations have been done using the MOPAC software.

# Chapter 5

# Conclusion

Recent studies by Gilson[8] and Grimme and co-workers[9] have used dispersion and hydrogen bonded corrected PM6 to compute the vibrational free energy contribution to the standard binding free energy for host-guest systems. However, computing this vibrational free energy contribution can be complicated by the presence of one or more imaginary frequencies in the vibrational analysis, as these frequencies will be ignored in thermodynamic calculations, and these numerical problems can introduce a significant error in the binding free energy.

In this thesis we address this problem by developing the PM6-D3H+ method and implementing it in the GAMESS program. The method combines the D3 dispersion correction developed by Grimme and co-workers with a modified version of the H+ hydrogen bond correction developed by Korth. The HF-3c method has recently also been used to calculate binding free energy[53] with good results. This method was also implemented in GAMESS and coupled with the fragmenting scheme of FMO to be able to have Hartree-Fock scale to protein sized systems. Overall, the accuracy of PM6-D3H+ and HF-3c is very similar to PM6-DH2 and PM6-DH+, with RMSD and MAD values within 0.1 kcal/mol of one another.

The ability to reproduce the hydrogen bond lengths of the S22 and S66 dataset was lacking in the MOPAC optimizers, but the optimization using GAMESS with the empirical corrected models, PM6-D3H+ and HF-3c the hydrogen bond lengths was close to MP2 with a RMSD of 0.08 Å. While the HF-3c method is much slower than PM6-D3H+, the ability to reproduce good hydrogen bonding without having specific hydrogen bond configuration terms in the model as with H+ is impressive. Geometry optimizations of the 88 complexes result in 82, 6, 0, and 0 geometries with 0, 1, 2, and $\geq 3$ imaginary frequencies using PM6-D3H+ implemented in GAMESS, and 56, 23, 6, 3 for HF-3c in GAMESS, while the corresponding numbers for PM6-DH+ implemented in MOPAC are 54, 17, 15, and 2. This decrease for the PM6 methods is mainly due to differences in geometry optimization algorithms and convergence criteria. Furthermore, the numerical stability of the method could be further increased by changing the definition of some of the dihedral angles used in the hydrogen bond correction term. However, this appears only to be an issue for very symmetric gasphase systems which is unlikely to occur in large heterogenous systems such as proteins.

The PM6-D3H+ method as implemented in GAMESS offers an attractive alternative to PM6-DH+ in MOPAC in cases where the LBFGS optimizer must be used and a vibrational analysis is needed, e.g. when computing vibrational free energies. While the GAMESS implementation is up to 10 times slower for geometry optimizations of proteins in bulk solvent, it is sufficiently fast to make geometry optimizations of small proteins practically feasible. The HF-3c method is many times slower than PM6, however the interface with FMO makes HF-3c able to scale to large system sizes, as the fragment scheme has an almost linear scaling with CPUs.

To further improve these methods the first step would to extend the PM6 method with *d*-integrals, primarily to include Sulfur in the calculations, as most protein structures has it. This requires some work as presently this code is not included in GAMESS, but some code from an old version of MNDOd by Thiel. PM6 with *d*-integrals is included in the source code of AMBER, which we have presently not worked with.

To give PM6 a chance to calculate the electronic structure of full-size proteins, an interface with FMO or similar methods is needed. Very recently an interface between FMO and the semi-empirical methods DFTB[54] has been made by Nishimoto *et al* and which definitely would be interesting to test out.

It would be interesting to see the 3c correction term from HF-3c be interface and parameterised with PM6, as the 3c corrections does not have any HB geometric correction and would therefor likely be better for geometry optimizations and vibrational analysis. The analytical gradients for the gCP and SRB term of 3c are not implemented, but should be possible to derive and implement without any troubles. The three-body gradient for the dispersion term is implemented only numerically in GAMESS, but the analytical is available which would be needed if gradient or hessian calculations are needed for this term. A NDDO based method with 3 corrections, all with analytical gradient would be a highly attractive method with great scalability.

The PM6-D3H+/PCM implementation in GAMESS has recently been used for protein structure refinement in order to predict reliable chemical shifts for the protein, as part of a Masters thesis.[55] Protein structures were refined using GAMESS, which then was used to calculate NMR data using Gaussian. The study shows, using hydrogen-bonded corrected schemes for structure refinement greatly affects the accuracy of the NMR calculation.

Work to reproduce the binding study of Gilson *et al*[8] with CB7 is currently under way, in collaboration with Hari Muddana and Mike Gilson at UC San Diego. The calculations seem much more reliable with much more stable vibrational analysis of the structures, however the calculations using GAMESS instead of MOPAC still needs some work as the results did not correlate well with experimental results for most of the ligands. The results of that study was out of the scope of this thesis as still a lot of work is needed to analyse the data and the need to introduce a new thermodynamic model similar to Gilson.

Work on enzyme reaction prediction using PM6 has been done before[56], but we are now working on a more qualitative approach to reproduce DFT level calculations on different sized reaction mechanism[57] with different semi-empirical methods. The idea of this project is also to have a similar test case for bio-chemical TS as with the BEGDB, and further calculation with more correlated method than DFT are needed, and thus the structures are made available online on github in hope other research group will re-optimize the structure and calculated the TS and product barriers with a method similar to CCSD(T)/CBS as with BEGDB.

# Bibliography

[1] J. Rezac and P. Hobza, “Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the ”Gold Standard”, CCSD(T) at the Complete Basis Set Limit?” *J. Chem. Theory Comput.*, vol. 9, p. 2151–2155, 2013, doi: 10.1021/ct400057w.

[2] J. J. P. Stewart, “Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements,” *J. Mol. Model*, vol. 13, pp. 1172–1213, 2007, doi: 10.1007/s00894-007-0233-4.

[3] M. Korth, M. Pitonák, J. Rezác, and P. Hobza, “A Transferable H-bonding Correction For Semiempirical Quantum-Chemical Methods,” *J. Chem. Theory Comput.*, vol. 6, p. 344–352, 2010, doi: 10.1021/ct900541n.

[4] J. Rezac and P. Hobza, “Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods,” *J. Chem. Theory Comput.*, vol. 8, p. 141–151, 2012, doi: 10.1021/ct200751e.

[5] M. Korth, “Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields,” *J. Chem. Theory Comput.*, vol. 12, pp. 33 803–3816, 2010, doi: 10.1021/ct100408b.

[6] N. D. Yilmazer and M. Korth, “Comparison of Molecular Mechanics, Semi-Empirical Quantum Mechanical, and Density Functional Theory Methods for Scoring Protein–Ligand Interactions,” *J. Phys. Chem.*, vol. 117, no. 27, pp. 8075–8084, 2013, doi: 10.1021/jp402719k.

[7] M. Korth and W. Thiel, “Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost As Accurate and Robust As DFT-GGA Methods for Organic Molecules,” *J. Chem. Theory and Comp.*, vol. 7, no. 9, pp. 2929–2936, 2011, doi: 10.1021/ct200434a.

[8] H. S. Muddana and M. K. Gilson, “Calculation of host-guest binding affinities using a quantum-mechanical energy model,” *J. Chem. Theory Comput.*, vol. 8, pp. 2023–2033, 2012, doi: 10.1021/ct3002738.

[9] S. Grimme, “Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory,” *Chem. Eur. J.*, vol. 18, no. 32, pp. 9955–9964, 2012, doi: 10.1002/chem.201200497.

[10] H. Muddana and M. K. Gilson, “Private communication,” 2013.

[11] J. C. Kromann, A. S. Christensen, C. Steinmann, M. Korth, and J. H. Jensen, “A third-generation dispersion and third-generation hydrogen bondingcorrected PM6 method: PM6-D3H+,” *PeerJ*, vol. 2, p. e449, 2014, doi: 10.7717/peerj.449.

[12] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J.A.Montgomery, “General atomic and molecular electronic structure system,” *J. Comput. Chem.*, vol. 14, pp. 1347–1363, 1993, doi: 10.1002/jcc.540141112.

[13] R. Sure and S. Grimme, “Corrected Small Basis Set Hartree-Fock Method for Large Systems,” *J. Comp. Chem.*, vol. 19, pp. 1672–85, 2013, doi: 10.1002/jcc.23317.

[14] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *J. Chem. Phys.*, vol. 132, p. 154104, 2010, doi: 10.1063/1.3382344.

[15] P. Jurecka, J. Cerný, P. Hobza, and D. Salahub, "Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations," *J Comput Chem.*, vol. 28, no. 2, pp. 555–69, 2007, doi: 10.1002/jcc.20570.

[16] S. P. A. Sauer, *Molecular Electromagnetism.* Oxford Graduate Texts, 2011.

[17] J. H. Jensen, *Molecular Modeling Basics.* CRC Press, 2010.

[18] A. Szabo and N. S. Ostlund, *Moderen Quantum Chemistry*, 1st ed. Macmillian Publishing Co., 1982.

[19] F. Jensen, *Introduction to Computational Chemistry*, 2nd ed. Wiley, 2007.

[20] J. A. Pople and D. L. Beveridge, *Approximate Molecular Orbital Theory.* McCraw-Hill, 1970.

[21] M. J. S. Dewar and W. Thiel, "A Semiempirical Model for the Two-Center Repulsion Integrals in the NDDO Approximation," *Theoret. Chim. Acta.*, vol. 46, pp. 89–104, 1977, doi: 10.1007/BF00548085.

[22] ——, "Ground states of molecules. 38. The MNDO method. Approximations and parameters," *J. A. Chem. Soc.*, vol. 15, p. 4899, 1977, doi: 10.1021/ja00457a004.

[23] W. Thiel and A. A. Voityuk, "Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group," *J. Phys. Chem.*, vol. 2, p. 616, 1996, doi: 10.1021/jp952148o.

[24] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, "Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model," *J. A. Chem. Soc.*, vol. 10, pp. 3902–3909, 1985, doi: 10.1021/ja00299a024.

[25] J. J. P. Stewart, "Optimization of Parameters for Semi-Empirical Methods I-Method," *J. Comp. Chem.*, vol. 10, pp. 209–220, 1989, doi: 10.1002/jcc.540100208.

[26] P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, 5th ed. Oxford University Press, 2011.

[27] JensenGroup, "Github: Third-generation hydrogen-bonding correction," 2014. [Online]. Available: https://github.com/jensengroup/hydrogen-bond-correction-f3

[28] M. Korth, "Empirical hydrogen-bond potential functions—an old hat reconditioned," *Chem. Phys. Chem.*, vol. 12, no. 17, pp. 3131–3142, 2011, doi: 10.1002/cphc.201100540.

[29] H. Kruse and S. Grimme, "A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems," *J. Chem. Phys.*, vol. 136, p. 154101, 2012, doi: 10.1063/1.3700154.

[30] J. R. P. Jurecka, K. E. Riley, J. Cerny, H. Valdes, K. Pluhackova, K. Berka, T. Rezac, M. Pitonak, J. Vondrasek, and P. Hobza, "Quantum chemical benchmark energy and geometry database a for molecular clusters and complex molecular systems (www.begdb.com): a users manual and examples," *Collect. Czech. Chem. Commun.*, vol. 73, pp. 1261–1270, 2008, doi: 10.1135/cccc20081261. [Online]. Available: http://www.begdb.com

[31] J. J. P. Stewart, "MOPAC2012," 2012, Stewart Computational Chemistry, Colorado Springs, CO, USA. [Online]. Available: http://openmopac.net

[32] D. G. Fedorov and K. Kitaura, "Extending the power of qantum chemistry to large systems with the fragment molecular orbital method," *J. Phys. Chem.*, vol. 111, pp. 6904–6914, 2007, doi: 10.1021/jp0716740.

[33] J. D. C. Maia, G. A. U. Carvalho, C. P. Mangueira, S. R. Santana, L. A. F. Cabral, and G. B. Rocha, "GPU Linear Algebra Libraries and GPGPU Programming for Accelerating MOPAC Semiempirical Quantum Chemistry Calculations," *J. Chem. Theory Comp.*, vol. 8, no. 9, pp. 3072–3081, 2012, doi: 10.1021/ct3004645.

[34] A. Klamt and G. Schuurmann, "COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient," *J. Chem. Soc., Perkin Trans.*, vol. 2, pp. 799–805, 1993, doi: 10.1039/P29930000799.

[35] P. Jurecka, J. Sponer, J. Cerny, and P. Hobza, "Benchmark database of accurate (mp2 and ccsd(t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs," *Phys Chem Chem Phys*, vol. 8, no. 17, pp. 1985–1993, 2006, doi: 10.1039/B600027D.

[36] J. Rezac, K. E. Riley, and P. Hobza, "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures," *J. Chem. Theory Comp.*, vol. 7, no. 8, pp. 2427–2438, 2011, doi: 10.1021/ct2002946.

[37] M. M. Francl and W. J. Pietro and W. J. Hehre and J. S. Binkley and M. S. Gordon and D. J. DeFrees and J. A. Pople, "Self-consistent molecular orbital methods. xxiii. a polarization-type basis set for second-row elements," *J. Chem. Phys.*, vol. 77, no. 7, pp. 3654–3665, 1982, doi: 10.1063/1.444267.

[38] Gordon, M.S.; Binkley, J.S.; Pople, J.A.; Pietro, W.J.; Hehre, W.J., "Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements," *J. Am. Chem. Soc.*, vol. 104, no. 10, pp. 2797–2803, 1982, doi: 10.1063/1.444267.

[39] P. C. Hariharan and J. A. Pople, "The influence of polarization functions on molecular orbital hydrogenation energies," *Theo. Chem. Acc.*, vol. 28, p. 213–222, 1973, doi: 10.1007/BF00533485.

[40] T. Nagata, K. Brorsen, D. G. Fedorov, K. Kazuo, and M. S. Gordon, "Fully analytic energy gradient in the fragment molecular orbital method," *J. Chem. Phys.*, vol. 134, no. 12, 2011, doi: 10.1063/1.3568010.

[41] R. Peverati and K. K. Baldridge, "Implementation and Performance of DFT-D with Respect to Basis Set and Functional for Study of Dispersion Interactions in Nanoscale Aromatic Hydrocarbons," *J. Chem. Theory Comput.*, vol. 12, pp. 2030–2048, 2008, doi: 10.1021/ct800252z.

[42] J. Tomansi, B. Mennucci, and R. Cammi, "Quantum mechanical continuum solvation models," *Chem. Rev.*, vol. 105, pp. 2999–3093, 2005.

[43] C. Steinmann, K. L. Blædel, A. S. Christensen, and J. H. Jensen, "Interface of the polarizable continuum model of solvation with semi-empirical methods in the games program," *PLoS ONE*, vol. 8, no. 7, p. e67725, 07 2013, doi: 10.1371/journal.pone.0067725.

[44] T. Nagata, D. G. Fedorov, H. Li, and K. Kitaura, "Analytic gradient for second order Møller-Plesset perturbation theory with the polarizable continuum model based on the fragment molecular orbital method," *J. Chem. Physics*, vol. 136, no. 20, 2012, doi: 10.1063/1.4714601.

[45] P. Su and H. Li, "Continuous and Smooth Potential Energy Surface for Conductor like Screening Solvation Model Using Fixed Points with Variable Areas," *J. Chem. Phys.*, vol. 130, p. 074109, 2009, doi: 10.1063/1.3077917.

[46] N. Kobko, L. Paraskevas, E. Rio, and J. J. Dannenberg, "Cooperativity in Amide Hydrogen Bonding Chains: Implications for Protein-Folding Models," *J. A. Chem. Soc.*, vol. 123, no. 18, pp. 4348–4349, 2001, doi: 10.1021/ja004271l.

[47] N. Kobko and J. J. Dannenberg, "Cooperativity in Amide Hydrogen Bonding Chains. Relation between Energy, Position, and H-Bond Chain Length in Peptide and Protein Folding Models," *J. Phys. Chem. A*, vol. 107, no. 48, pp. 10 389–10 395, 2003, doi: 10.1021/jp0365209.

[48] P. Jurecka, J. Sponer, J. Cerny, and P. Hobza, “Benchmark database of accurate (mp2 and ccsd(t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs,” *Phys. Chem. Chem. Phys.*, vol. 8, pp. 1985–1993, 2006, doi: 10.1039/B600027D.

[49] W. Kabsch, “A solution for the best rotation to relate two sets of vectors,” *Acta Cryst.*, vol. A32, pp. 922–923, 1976, doi: 10.1107/S0567739476001873.

[50] J. C. Kromann, “GitHub: Calculate RMSD from two XYZ structures,” 2013. [Online]. Available: https://github.com/charnley/rmsd

[51] JensenGroup, “GitHub: Optimized Protein Structures,” 2014. [Online]. Available: https://github.com/jensengroup/optimized-protein-structures

[52] Schrodinger, LLC, “Pymol: The pymol molecular graphics system, schrodinger, llc.” [Online]. Available: http://www.pymol.org

[53] R. Sure, J. Antony, and S. Grimme, “Blind prediction of binding affinities for charged supramolecular host–guest systems: Achievements and shortcomings of dft-d3,” *J. Phys. Chem.*, vol. 118, no. 12, pp. 3431–3440, 2014, doi: 10.1021/jp411616b.

[54] Y. Nishimoto, D. G. Fedorov, and S. Irle, “Density-functional tight-binding combined with the fragment molecular orbital method,” *J. Chem. Theory Comp.*, 2014, doi: 10.1021/ct500489d.

[55] A. Larsen, “Protein chemical shift prediction,” *arXiv:1409.6772 [physics.chem-ph]*, 2014. [Online]. Available: http://arxiv.org/abs/1409.6772

[56] M. R. Hediger, C. Steinmann, L. D. Vico, and J. H. Jensen, “A computational method for the systematic screening of reaction barriers in enzymes: searching for bacillus circulans xylanase mutants with greater activity towards a synthetic substrate,” *PeerJ*, p. 1:e111, 2013, doi: 10.7717/peerj.111.

[57] JensenGroup, “Reactions barriers of various proteins,” 2014. [Online]. Available: https://github.com/jensengroup/reaction-barriers-proteins

# Supplementary Information

- S.1 S22 and S66 Complexes optimized with OPTTOL = 0.0001
- S.2 S22 and S66 Complexes with imaginary frequencies
- S.3 Selected Complexes from JSCH-2005
- S.4 GAMESS Header examples

## S.1 S22 and S66 Complexes optimized with OPTTOL = 0.0001

**Table S1**

| Set | ID | Name |
|---|---|---|
| s22 | 08 | Methanedimer |
| s22 | 12 | Pyrazinedimer |
| s22 | 18 | Benzeneammoniacomplex |
| s66 | 04 | WaterPeptide |
| s66 | 23 | AcNH2Uracil[a] |
| s66 | 33 | PyridineEthene |
| s66 | 37 | CyclopentaneNeopentane |
| s66 | 38 | CyclopentaneCyclopentane |
| s66 | 57 | BenzenePeptideNHpi |
| s66 | 65 | PyridineEthyne[a] |
| s66 | 66 | MeNH2Pyridine |

[a] Necessary to also set ihrep to 20.

## S.2 S22 and S66 Complexes with imaginary frequencies

**Table S2**

| Set | ID | Name | No. *i*-freq |
|---|---|---|---|
| | | PM6[a] | |
| s66 | 30 | BenzeneEthene | 1 |
| s66 | 65 | PyridineEthyne | 1 |
| | | PM6-D3H+[a] | |
| s22 | 01 | Ammoniadimer | 1 |
| s22 | 15 | Adeninethyminecomplexstack | 1 |
| s22 | 20 | BenzenedimerTshaped | 1 |
| s66 | 16 | PeptideWater | 1 |
| s66 | 30 | BenzeneEthene | 1 |
| s66 | 42 | UracilCyclopentane | 1 |
| | | HF-3c[a] | |
| s22 | 01 | Ammoniadimer | 1 |
| s22 | 08 | Methanedimer | 3 |
| s22 | 10 | BenzeneMethanecomplex | 3 |
| s22 | 11 | Benzenedimerparalleldisplacer | 2 |
| s22 | 14 | Indolebenzenecomplexstack | 1 |
| s22 | 16 | Etheneethynecomplex | 1 |
| s22 | 18 | Benzeneammoniacomplex | 1 |
| s22 | 19 | BenzeneHCNcomplex | 2 |
| s22 | 20 | BenzenedimerTshaped | 3 |
| s22 | 21 | IndolebenzeneTshapecomplex | 2 |
| s66 | 04 | WaterPeptide | 1 |
| s66 | 05 | MeOHMeOH | 1 |
| s66 | 06 | MeOHMeNH2 | 1 |
| s66 | 09 | MeNH2MeOH | 1 |
| s66 | 13 | PeptideMeOH | 1 |
| s66 | 14 | PeptideMeNH2 | 1 |
| s66 | 19 | MeOHPyridine | 1 |
| s66 | 21 | AcNH2AcNH2 | 1 |
| s66 | 23 | AcNH2Uracil | 1 |
| s66 | 24 | BenzeneBenzenepipi | 2 |
| s66 | 27 | BenzenePyridinepipi | 1 |
| s66 | 28 | BenzeneUracilpipi | 1 |
| s66 | 30 | BenzeneEthene | 1 |
| s66 | 33 | PyridineEthene | 1 |
| s66 | 44 | EthenePentane | 1 |
| s66 | 47 | BenzeneBenzeneTS | 1 |
| s66 | 48 | PyridinePyridineTS | 1 |
| s66 | 49 | BenzenePyridineTS | 2 |
| s66 | 50 | BenzeneEthyneCHpi | 2 |
| s66 | 53 | BenzeneAcNH2NHpi | 1 |
| s66 | 54 | BenzeneWaterOHpi | 1 |
| s66 | 57 | BenzenePeptideNHpi | 1 |
| s66 | 66 | MeNH2Pyridine | 1 |

[a] The calculations have been done using the GAMESS software.

[b] The calculations have been done using the MOPAC software.

**Table S3**

| Set | ID | Name | No. $i$-freq |
|---|---|---|---|
| | | PM6[b] | |
| s22 | 01 | Ammoniadimer | 2 |
| s22 | 04 | Formamidedimer | 2 |
| s22 | 05 | Uracildimerhbonded | 2 |
| s22 | 06 | 2pyridoxine2aminopyridinecomplex | 3 |
| s22 | 07 | AdeninethymineWatsonCrickcomplex | 2 |
| s22 | 11 | Benzenedimerparalleldisplaced | 1 |
| s22 | 14 | Indolebenzenecomplexstack | 1 |
| s22 | 19 | BenzeneHCNcomplex | 2 |
| s22 | 20 | BenzenedimerTshaped | 3 |
| s66 | 04 | WaterPeptide | 1 |
| s66 | 05 | MeOHMeOH | 1 |
| s66 | 06 | MeOHMeNH2 | 1 |
| s66 | 08 | MeOHWater | 1 |
| s66 | 09 | MeNH2MeOH | 1 |
| s66 | 10 | MeNH2MeNH2 | 1 |
| s66 | 14 | PeptideMeNH2 | 2 |
| s66 | 16 | PeptideWater | 1 |
| s66 | 17 | UracilUracilBP | 2 |
| s66 | 18 | WaterPyridine | 1 |
| s66 | 20 | AcOHAcOH | 1 |
| s66 | 22 | AcOHUracil | 1 |
| s66 | 23 | AcNH2Uracil | 2 |
| s66 | 25 | PyridinePyridinepipi | 1 |
| s66 | 28 | BenzeneUracilpipi | 1 |
| s66 | 35 | NeopentanePentane | 1 |
| s66 | 36 | NeopentaneNeopentane | 6 |
| s66 | 37 | CyclopentaneNeopentane | 1 |
| s66 | 39 | BenzeneCyclopentane | 1 |
| s66 | 41 | UracilPentane | 1 |
| s66 | 42 | UracilCyclopentane | 2 |
| s66 | 44 | EthenePentane | 1 |
| s66 | 47 | BenzeneBenzeneTS | 2 |
| s66 | 48 | PyridinePyridineTS | 1 |
| s66 | 49 | BenzenePyridineTS | 1 |
| s66 | 59 | EthyneWaterCHO | 2 |
| s66 | 63 | BenzeneAcOH | 1 |
| s66 | 66 | MeNH2Pyridine | 3 |

[a] The calculations have been done using the GAMESS software.
[b] The calculations have been done using the MOPAC software.

**Table S4**

| Set | ID | Name | No. *i*-freq |
|---|---|---|---|
| | | PM6-DH+[b] | |
| s22 | 01 | Ammoniadimer | 2 |
| s22 | 05 | Uracildimerhbonded | 1 |
| s22 | 06 | 2pyridoxine2aminopyridinecomplex | 2 |
| s22 | 07 | AdeninethymineWatsonCrickcomplex | 1 |
| s22 | 10 | BenzeneMethanecomplex | 3 |
| s22 | 18 | Benzeneammoniacomplex | 2 |
| s22 | 19 | BenzeneHCNcomplex | 2 |
| s22 | 20 | BenzenedimerTshaped | 2 |
| s22 | 21 | IndolebenzeneTshapecomplex | 1 |
| s66 | 08 | MeOHWater | 2 |
| s66 | 10 | MeNH2MeNH2 | 1 |
| s66 | 12 | MeNH2Water | 1 |
| s66 | 13 | PeptideMeOH | 1 |
| s66 | 14 | PeptideMeNH2 | 1 |
| s66 | 15 | PeptidePeptide | 1 |
| s66 | 16 | PeptideWater | 2 |
| s66 | 17 | UracilUracilBP | 2 |
| s66 | 19 | MeOHPyridine | 1 |
| s66 | 20 | AcOHAcOH | 2 |
| s66 | 22 | AcOHUracil | 1 |
| s66 | 23 | AcNH2Uracil | 2 |
| s66 | 24 | BenzeneBenzenepipi | 1 |
| s66 | 25 | PyridinePyridinepipi | 2 |
| s66 | 36 | NeopentaneNeopentane | 9 |
| s66 | 42 | UracilCyclopentane | 2 |
| s66 | 45 | EthynePentane | 1 |
| s66 | 46 | PeptidePentane | 1 |
| s66 | 47 | BenzeneBenzeneTS | 1 |
| s66 | 48 | PyridinePyridineTS | 1 |
| s66 | 49 | BenzenePyridineTS | 2 |
| s66 | 52 | BenzeneAcOHOHpi | 2 |
| s66 | 55 | BenzeneMeOHOHpi | 1 |
| s66 | 66 | MeNH2Pyridine | 1 |

[a] The calculations have been done using the GAMESS software.
[b] The calculations have been done using the MOPAC software.

## S.3 Selected Complexes from JSCH-2005

**Table S5**

| BEGDB ID | Name |
|---|---|
| 1018 | G...U wobble |
| 1017 | I...C WC |
| 1020 | U...U |
| 1021 | U...U pl |
| 1084 | A...T S1 |
| 1014 | A...T WC |
| 1082 | G...C S |
| 1012 | G...C WC(1) |
| 1015 | mA...mT H |
| 1085 | mA...mT S |
| 1083 | mG...mC S |
| 1013 | mG...mC WC |

## S.4 GAMESS Header examples

### PM6-D3H+ Optimization and vibrational analysis

```
$basis
   gbasis=PM6-D3H+   ! Use the PM6 method w/ D3 and H+ correction
$end

$contrl
   scftyp=RHF        ! Use Restricted Hartree-fock
   icharg=0          ! Total molecule charge
   runtyp=optimize   ! Do a geometry optimization
$end

$scf
   npunch=1          ! less output during SCF iterations
$end

$statpt
   opttol=5.0e-4     ! convergence critria
   nstep=500         ! Maximum no. of steps

   hssend=.T.        ! do hessian calculation after optimization
$end

$force
   nvib=2            ! force calculation using centered finite difference
       scheme
   method=seminum    ! Use semi-numerical scheme for force calculation
$end
```

### PM6-D3H+/PCM Optimization and vibrational analysis

```
$basis
   gbasis=PM6-D3H+   ! Use the PM6 method w/ D3 and H+ correction
$end

$contrl
   scftyp=RHF        ! Use Restricted Hartree-fock
   icharg=0          ! Total molecule charge
   runtyp=optimize   ! Do a geometry optimization
$end

$scf
   npunch=1          ! less output during SCF iterations
$end

$statpt
   opttol=5.0e-4     ! convergence critria
   nstep=500         ! Maximum no. of steps

   hssend=.T.        ! do hessian calculation after optimization
$end

$force
   nvib=2            ! force calculation using centered finite difference scheme
   method=seminum    ! Use semi-numerical scheme for force calculation
$end

! Solvent settings
$pcm
   solvnt=WATER
   mxts=15000        ! The maximum number of tesserae
$end

$tescav
   mthall=4          ! Use the FIXPVA scheme
   ntsall=60         ! The density of tesserae
$end
```

## PM6-D3H+/PCM Optimization GAMESS header w/ convergence help

```
$basis
    gbasis=PM6-D3H+   ! Use the PM6 method w/ D3 and H+ correction
$end

$contrl
    scftyp=RHF        ! Use Restricted Hartree-fock
    icharg=0          ! Total molecule charge
    runtyp=optimize   ! Do a geometry optimization
$end

$scf
    npunch=1          ! less output during SCF iterations
$end

$statpt
    opttol=1.0e-4     ! convergence critria
    nstep=500         ! Maximum no. of steps

    hssend=.T.        ! do hessian calculation after optimization

    ihrep=20          ! Update Hessian every nth step
    projct=.F.        ! flag to eliminate translation and rotational degress of freedom
$end

$force
    nvib=2            ! force calculation using centered finite difference scheme
    method=seminum    ! Use semi-numerical scheme for force calculation
$end

! Solvent settings
$pcm
    solvnt=WATER
    mxts=15000        ! The maximum number of tesserae
$end

$tescav
    mthall=4          ! Use the FIXPVA scheme
    ntsall=60         ! The density of tesserae
$end
```