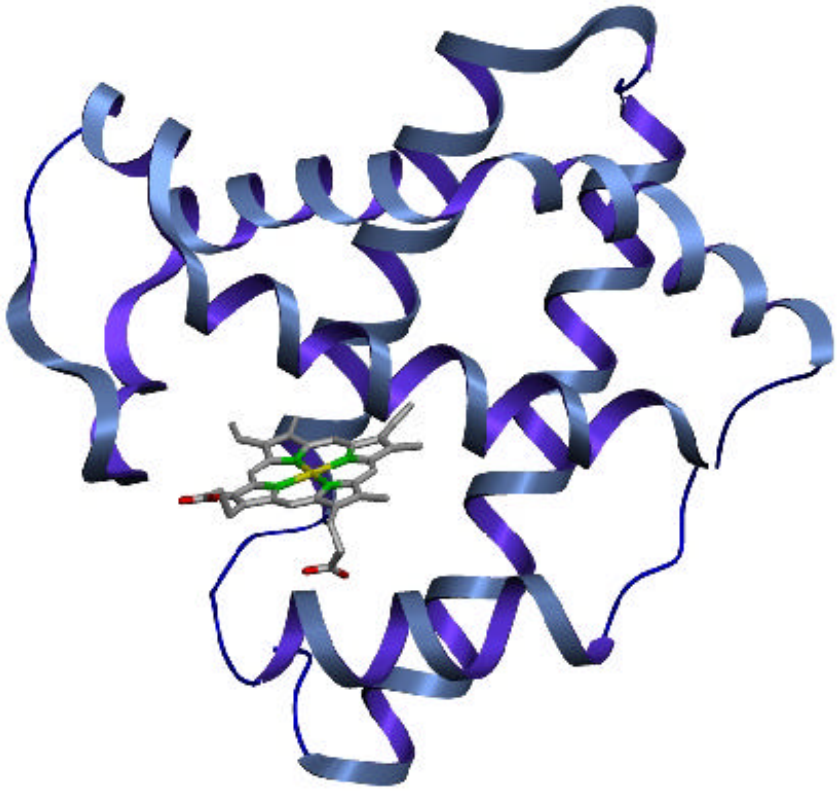
An Investigation of the Reactivity of Myoglobin Using Quantum Mechanical Techniques



A Fourth Year Chemistry Research Project Report
June 2000

Supervisor: Dr. I. Gould

Ross Walker
(96Chem518)
M.Sci. Chemistry(F103)

Imperial College of Science
Technology & Medicine

Table of Contents

Abstract ***5***

1. Introduction & Background ***6***

1.1 Proteins
1.2 Myoglobin - Background
1.3 Myoglobin - General Structure and Reactivity
1.4 Myoglobin - Biological Significance
1.5 Myoglobin - Ligand Discrimination

2. Background Theory & Methodology ***12***

2.1 Background Theory
2.2 The Hartree-Fock Method
2.3 Basis Sets
2.4 Electron Correlation
2.5 Density Functional Theory
2.6 Fast Multipole Methods
2.7 Binding Energy Calculations

3. Initial Scaling & Feasibility Study ***38***

3.1 FMM Scaling Tests

4. Development of Working Methodology for Study ***44***

4.1 Level of Theory
4.2 Selection of DFT Functional
4.3 Optimisation of Level Shifting
4.4 Optimisation of Annealing Steps
4.5 Frequency of Incremental Fock Builds
4.6 Selection of Spin State
4.7 Basis Set Selection
4.8 Summary

5. Ligand Binding Investigations ***60***

5.1 Control System (Iron Protoporphyrin IX)
5.2 Carbon Monoxide Binding to Heme Systems
5.3 Abandoned Calculations

6. Conclusions and Future Work ***76***

Appendix A - System Benchmark Calculations ***78***

A.1 44 Processor Silicon Graphics ONYX 2
A.2 Dual Intel Celeron
A.3 Benchmark Calculations

Cover Illustration[1]*:* Image showing the heme unit and tertiary structure of $P2_1$ Fe(III)-aquo myoglobin (sperm whale) at 1.0 Å.

Appendix B - Custom Software Developed For This Research ***85***

B.1 Computational Chemistry Tools V0.6
B.2 Output Conversion Module
B.3 Text File Character Extraction
B.4 Global Search and Replace
B.5 Time Conversion Module
B.6 Unit Conversion

Appendix C – Modification to Gaussian Code (ONIOM) ***90***

C.1 Gaussian Source Code Modification

Appendix D – Project Support CDRom ***93***

References ***95***

Acknowledgements

"The underlying laws necessary for the mathematical theory of large parts of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."

(Paul Dirac - 1929)[2]

Abstract

A working methodology that solves the convergence problems associated with density functional calculations of iron containing systems has been developed and successfully employed in a study of carbon monoxide and oxygen binding to myoglobin derivatives.

Calculations show that carbon monoxide binds to porphyrin systems with an affinity 30,000 times greater than oxygen. Studies of carbon monoxyheme and carbon monoxyheme+his93 show that the proximal histidine (his 93) plays a very small part in reducing the CO binding energy (15 % reduction in binding energy) and that linear CO is the preferred geometry for both deoxyheme and deoxyheme+his93. Calculations on oxygen binding while unfinished due to time restraints imply that a bent geometry is preferable in all cases and that the binding energy would seem to be increased by the proximal histidine.

1. Introduction & Background

1.1 Proteins

Proteins are a class of compounds, into which myoglobin falls, that are, in the form of enzymes, responsible for catalysis and control of reactions in living organisms. They are essentially complex polymers made up from repeating simple units called amino acids[3]. This complexity arises from the fact that whereas most polymers are made up of one or two repeating units called monomers, with proteins there are essentially 20 different monomers (amino acids) that are linked together in a very specific way.

The 20 different amino acids that form proteins all have their own individual chemical properties. However, when combined into a single molecule these properties result in a protein that has a reactivity more specific than possible with standard organic molecules[4]. It is this specificity and the huge diversity of available proteins that makes life on earth possible.

Proteins are responsible for all of the reactions needed to sustain life from the transport and storage of oxygen in mammals (Haemoglobin/Myoglobin) to Vertebrate movement (Myosin) to the replication of DNA (DNA Polymerase). Proteins can catalyse reactions that scientists using current technology, cannot even hope to replicate in the laboratory. A prime example is the protein nitrogenase which can successfully convert molecular nitrogen into ammonia at room temperature and pressure. *cf.* the Haber process, which is the current industrial method for the production of ammonia from nitrogen and hydrogen, that, even with a catalyst present, requires temperatures in excess of 400^{o}C and pressures greater than 200 atmospheres to give a yield of less than 40 % ammonia[5].

The ability of proteins to exert precise control over reaction kinetics and stereochemistry makes their study a very interesting and potentially rewarding area. Consequently the ability to model and understand protein function is high on the list of modern scientific objectives. In the short term such studies would have a range of benefits from the designing of medicinal drugs to the understanding of disease processes. A number of pharmaceuticals work by inhibiting proteins, such as HIV protease inhibitors, thus further protein research will allow advances to be made in the effectiveness of these type of drugs. There is also currently a lot of interest in combating cancer by inhibiting specific proteins within a cancer cell.

Another example of where proteins can be of interest to medicine comes from the protein myoglobin which is responsible for the storage of oxygen within muscle. Recent research[6] has shown that myoglobin can act as a marker for cardiovascular problems in humans. Hence being able to understand the way in which myoglobin functions could help in the construction of assays to predict heart disease and related problems.

In the long term, understanding precisely how proteins function may make it possible to design artificial proteins to carry out reactions that are currently impossible using traditional techniques. This field is known as protein engineering and is a very active and controversial research field at present.

1.2 Myoglobin - Background

Myoglobin pictured opposite (*figure 1.1*) is an iron-based oxygen transport and storage protein found in the muscles of vertebrates.

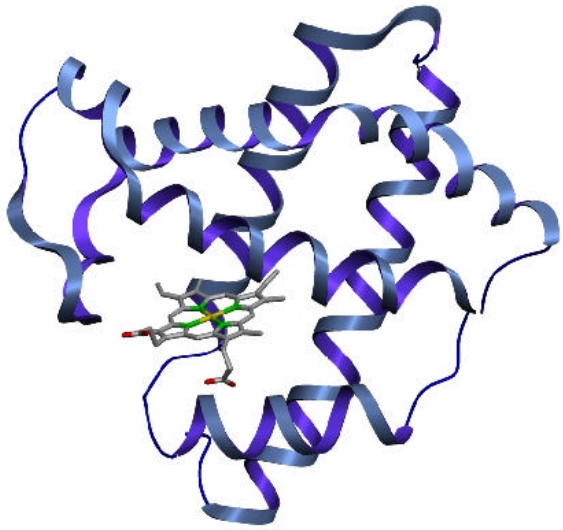
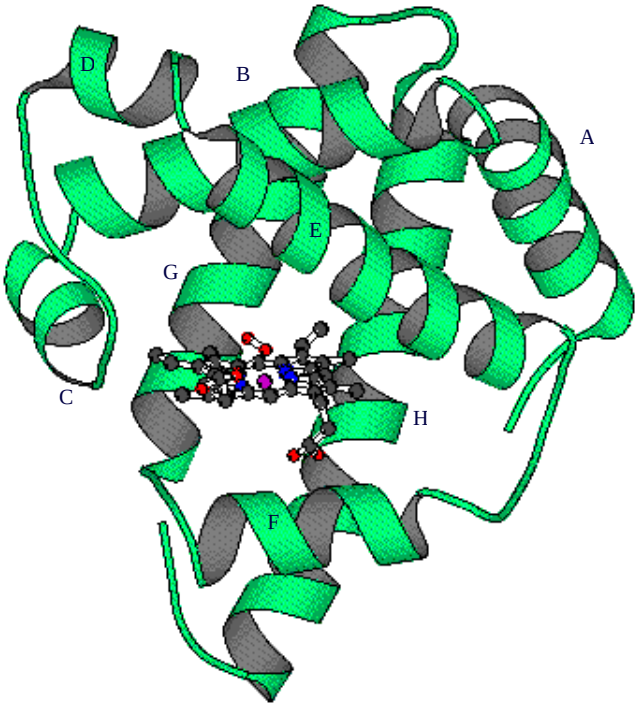


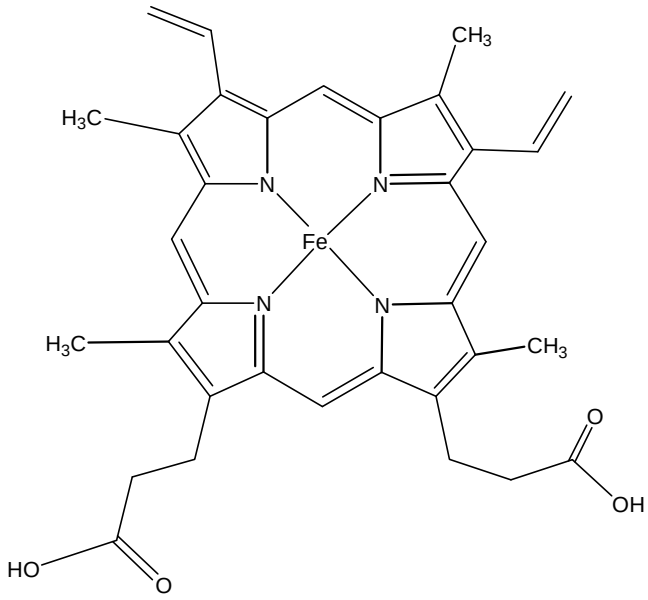
Figure 1.1: $P2_1$ Fe(III)-aquo myoglobin (sperm whale) at 1.0 Å. Adapted from Maurus, R. et al. *Biophys. Bioch. Acta.* 1341, **1997**.

Myoglobin was one of the first protein structures to be solved using x-ray crystallography by John Kendrew and co-workers in 1958[7]. Since then a large number of studies have been undertaken and its structure has now been solved using both neutron and X-ray diffraction to a resolution of 1 angstrom *(PDB 1A6M)*. Even at this high resolution, however, there is still debate over the exact structure of the ligand binding site which is composed of a single heme unit. It is the structure of this binding site that has been investigated in this project.

1.3 Myoglobin - General Structure and Reactivity

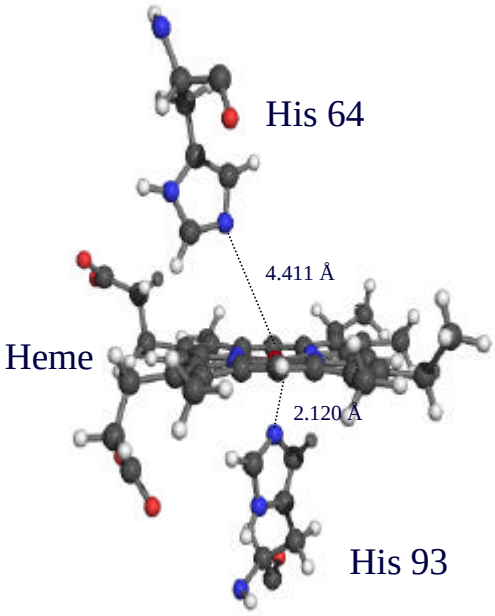
Myoglobin is a compact, predominantly (75 %) alpha helical, globular protein consisting of only 153 amino acid residues which makes it sufficiently small to be accurately studied using theoretical techniques and ideal for the development and trial of new methodologies for protein study. The 8 alpha helices, labelled A to H in figure 1.2a form an amphipathic pocket that stabilises the essential prosthetic group, iron protoporphyrin IX (*figure 1.2b*). The latter is referred to as heme when in the Fe^{2+} oxidation state and hemin when in the Fe^{3+} oxidation state.





Figures 1.2a & 1.2b: Illustration of the 8 alpha helices of myoglobin and a schematic of the heme unit (Iron protoporphyrin 9).

The iron atom is coordinated to the protein via the proximal histidine residue (his 93) and to the protoporphyrin ring by the four pyrrole nitrogen atoms (*figure 1.3*). There is also a second histidine present, the distal histidine (his 64), at a distance of 4.4 Å from the iron atom.



Figures 1.3: Image showing the location of the proximal (his 93) and distal (his 64) histidine units with respect to the heme unit.

Myoglobin binds ligands through the one remaining iron coordination position on the distal face of the heme. The iron atom can exist in two physiologically relevant oxidation states. In the oxidised Fe^{3+} (ferric) state myoglobin can bind a water molecule or a number of different anions including N_3^-, CN^-, NO_2^-, SCN^- and F^-. In the Fe^{2+} (ferrous) state the iron can bind oxygen, carbon monoxide and other neutral ligands including nitrogen monoxide, aryl nitroso compounds and alkyl iso-cyanides. This research has concentrated on the binding of oxygen and carbon monoxide to myoglobin in the ferrous state as these two ligands are the most physiologically interesting.

1.4 Myoglobin - Biological Significance

Hemoglobin and Myoglobin act as partners in the transport and storage of oxygen in vertebrates. While hemoglobin is found in high concentrations in red blood cells and is responsible for the uptake of oxygen in the lungs and subsequent transport around the body it is myoglobin which stores this oxygen in the muscle until the body requires it. This is illustrated, in a somewhat simplified view, by figure 1.4.

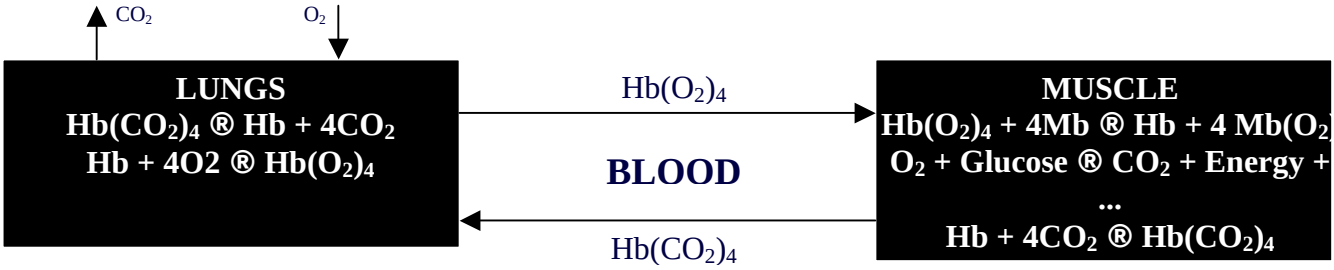


Figure 1.4: Schematic of the roles of hemoglobin and myoglobin in respiration.

Oxygen is absorbed through the lungs where it binds to hemoglobin and is then transported via the bloodstream to the muscles. Here, due to a difference in binding affinities, it is released from the hemoglobin and taken up by myoglobin which then acts as an oxygen store until the muscle requires it for respiration. Waste carbon dioxide is then transported back to the lungs where it is expelled.

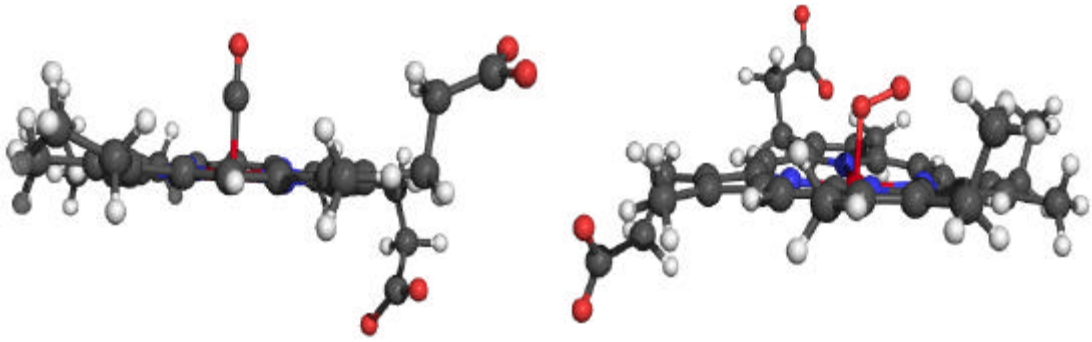
1.5 Myoglobin - Ligand Discrimination

Unhindered 5 coordinate heme binds carbon monoxide 30,000 to 100,000 times more strongly than oxygen[8]. However, when the heme is embedded in the protein matrix of myoglobin the ratio is reduced a thousand fold to between 30 and 100 times. Thus myoglobin can selectively discriminate between carbon monoxide and oxygen.

This mechanism for preferentially binding oxygen has evolved to prevent the inhibition of oxygen transport and storage from endogenously produced carbon monoxide. Both neurotransmission activity and the catabolic breakdown of heme would produce sufficient carbon monoxide to inhibit myoglobin's function if the protein had the same relative oxygen and carbon monoxide affinities as bare heme. The question that arises therefore is how is this discrimination achieved?

There are currently two schools of thought on how myoglobin discriminates between carbon monoxide and oxygen. The first (*§ 1.5.1*) centres around the idea that while CO binds to bare heme in a linear configuration (*figure 1.5a*) when binding to heme in myoglobin it is forced to bind in an unfavourable bent configuration. This disrupts the π back bonding and so destabilises the complex

and reduces the binding energy. Conversely oxygen binds to both heme and myoglobin in a bent configuration (*figure 1.5b*). The second argument (*§ 1.5.2*) is that the oxygen bound state is stabilised via the formation of favourable hydrogen bonds.



CO O_2

Figures 1.5a & 1.5b: Illustration of the binding geometries of CO (a) and O_2 (b) to bare heme.

1.5.1 Discrimination via Carbon Monoxide Destabilisation

As mentioned above one argument is that selectivity is achieved by destabilisation of carbon monoxide binding by disruption of the pi back bonding. The mechanism by which this is achieved, however, is still in dispute. The first high resolution X-ray structure of carbon monoxy myoglobin showed the Fe-C-O geometry to be bent and tilted away from the distal side chain (Brookhaven Protein Data Bank 1MBC)[9]. This was interpreted in terms of a steric repulsion between the distal histidine and the CO ligand[10,11,12]. The bend angles observed in the crystal structures, however, are large and it is therefore difficult to account for such large strain energies being exerted by the distal side chain[13].

It should also be noted that there are large variations in the CO geometries predicted by X-ray and neutron diffraction studies with structures showing variously distorted (tilted and / or bent) CO conformations[14,15,16,17]. Thus despite extensive high-resolution crystallographic data on myoglobin it is not yet possible from purely structural data to determine whether the Fe-C-O unit is linear but tilted away from the heme normal or whether it is bent with the Fe-C perpendicular to the heme plane and the C-O bond off axis[18]. The destabilisation of carbon monoxide on the grounds of pure steric interactions with the distal histidine is therefore disputed.

Another theory to account for CO destabilisation proposed by P. Jewsbury *et al.*[19] on the basis of *ab initio* calculations is that it is the proximal histidine that affects the binding affinity of carbon

monoxide by inhibiting movement of the iron atom relative to the plane of the porphyrin ring. This study relied on a large number of approximations in the calculations, not least of all locking the heme plane, distal histidine and proximal histidine in C_s symmetry. These massive approximations throw doubt on the results and so one of the goals of this research has been to examine the role of the proximal histidine theoretically in a more rigorous manner.

1.5.2 Discrimination via Oxygen Stabilisation

A second argument for the ligand discrimination is that the oxygen binding affinity is increased due to favourable interactions (a hydrogen bond) with the distal histidine (*figure 1.6*). The presence of such a hydrogen bond was first proposed by Pauling[20] and later confirmed by neutron diffraction studies of oxymyoglobin[21,22]. The quantitative effect that such a hydrogen bond has on the binding affinity of oxygen is still unclear. Thus a second objective of this research has been to investigate, using *ab initio* methods, whether such a hydrogen bond is present and to quantify its effect on the binding affinity of oxygen to myoglobin.

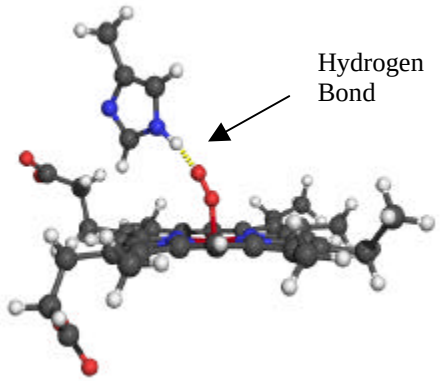


Figure 1.6: Illustration showing possible location of stabilising hydrogen bond to oxygen at the myoglobin active site.

1.5.3 This Work

The true reason for discrimination is probably a combination of both oxygen stabilisation and carbon monoxide destabilisation. Thus the main aim of this project has been to investigate, using a quantum mechanical approach, the role of the proximal and distal histidine residues in influencing the relative binding affinities of carbon monoxide and oxygen to the myoglobin active site.

2. Background Theory & Methodology

A number of different theoretical methods have been used in the course of this investigation a brief description of the theory of which will be given here.

2.1 Background Theory[a]

The Schrödinger equation is a well known entity in modern science. In its barest, time independent form it is represented by the eigenvalue relationship[23].

$$H\Psi = E\Psi \qquad 2.1$$

In the above equation H is the short hand notation for the Hamiltonian operator which operates on the mathematical function Ψ, which represents the wavefunction describing the system, to yield the energy E. Writing this equation in the form given in equation 2.1 disguises the fact that this is actually a set of differential equations each with a function Ψ_n corresponding to each allowed energy E_n.

Once the wave function is known for a particular system state it is then possible, in theory, to determine any physical observable using equation 2.2[24]

$$\text{Observable} = \frac{\int \Psi^* \langle \text{operator} \rangle \Psi \mathrm{d}t}{\int \Psi^* \Psi \mathrm{d}t} \qquad 2.2$$

where the operator used is that appropriate to the observable required. i.e. the Hamiltonian H for energy, another for dipole moment, charge density etc.

2.1.1 The Hamiltonian Operator

For a time-independent, multi-electron system the molecular Hamiltonian is[25]

$$H = -\frac{\hbar^2}{2m_e}\sum_i^n \nabla_i^2 - \frac{\hbar^2}{2}\sum_a^N \frac{1}{m_a}\nabla_a^2 - \sum_a^N \sum_i^n \frac{Z_a e^2}{4\pi e_0 r_{ai}} + \sum_a^N \sum_{b>a}^N \frac{Z_a Z_b e^2}{4\pi e_0 r_{ab}} + \sum_j^n \sum_{i>j}^n \frac{e^2}{4\pi e_0 r_{ij}} \qquad 2.3$$

[a] Presented here is a brief overview of the background theory underpinning the methods used for calculating electronic structure. For a more in-depth discussion of quantum mechanics the reader is referred to P. W. Atkins & R. S. Friedman, "Molecular Quantum Mechanics". Chapter 9 of this book also gives a well written discussion of the methods behind calculating electronic structure.

where *n* represents the number of electrons and *N* the number of nuclei. The first term is the operator for the electron kinetic energy (T_e). Similarly the second term is the operator for the nuclei kinetic energy (T_n). The third term takes account of the potential energy arising from electron-nuclei attraction (V_{ne}). The fourth term represents the potential energy due to repulsion between nuclei (V_{nn}) and the final term represents the electron-electron repulsions (V_{ee}). The total Hamiltonian operator can thus be represented by a summation of the kinetic and potential energies of the nuclei and electrons (*eq. 2.4*).

$$H_{tot} = T_e + T_n + V_{ne} + V_{nn} + V_{ee} \qquad 2.4$$

2.1.2 The Born-Oppenheimer Approximation[26]

Oppenheimer and Born showed, in 1927, that equation 2.4 could be greatly simplified by exploiting the fact that nuclei have a much greater mass than electrons and hence can be considered to remain fixed whilst the electrons move. This allows the nuclear kinetic energy (T_n) to be neglected and the nuclear potential energy (V_{nn}) to be considered constant and so treated as a separate entity and added to the total energy at the end of a calculation.

Thus to find the energy of a system, in the Born Oppenheimer approximation, it becomes necessary to find only the pure electronic wavefunction given by the modified form of the Schrödinger equation (*eq. 2.5*).

$$H_{el}\Psi_{el} = E_{el}\Psi_{el} \qquad 2.5$$

where H_{el} is the electronic Hamiltonian (*eq. 2.6*)

$$H_{el} = -\frac{\hbar^2}{2m_e}\sum_i^n \nabla_i^2 - \sum_a^N \sum_i^n \frac{Z_a e^2}{4\pi e_0 r_{ai}} + \sum_j^n \sum_{i>j}^n \frac{e^2}{4\pi e_0 r_{ij}} \qquad 2.6$$

Solution of equation 2.5 then yields the electronic energy E_{el} which when added to the nuclear-nuclear potential (V_{nn}) gives the total energy for the system (*eq. 2.7*).

$$E_{tot} = E_{el} + V_{nn} \qquad 2.7$$

Thus the problem has been reduced to finding two terms, the electronic energy and the nuclear repulsion energy. However, this is not as simple as might at first be thought as solution of the

purely electronic Schrödinger equation (*eq. 2.5*) is not possible for systems with more than one electron. The reason for this lies with the electron-electron repulsion term in the Hamiltonian. This is a simple coulombic repulsion term, however, the value of this term depends on the distance between each of the electrons which unfortunately cannot be known prior to the calculation. e.g. In a three electron system the position of electron 1 depends on the positions of electrons 2 and 3 while the position of electron 2 depends on that of 1 and 3. Thus one has to use further approximations and iterative methods to find the wavefunction for a many electron system.

2.2 The Hartree-Fock Method

In 1928 Hartree devised a method that made it possible to find, to a reasonable approximation, the most accurate wavefunction, and hence solution to the Schrödinger equation for a poly electronic system[27]. The approximation employed by Hartree was to treat the electron-electron repulsions in an average way. In this approximation each electron moves in a field of nuclei and the average field of the other n-1 electrons. This approximation allows the many electron wavefunction to be replaced by a product of one-electron wavefunctions (*eq. 2.8*).

$$\Psi(r) = \Psi_1(r_1)\Psi_2(r_2).....\Psi_n(r_n) \qquad 2.8$$

where $\Psi_i(r_i)$ is a spatial orbital that is a function of the position vector ***r*** such that the probability of finding an electron at a distance $d\boldsymbol{r}$ from ***r*** is $|\Psi_i(r)|^2 dr$.[28] The overall wavefunction therefore depends on all the electron co-ordinates and parametrically on the nuclear locations. Thus different nuclear arrangements will produce different solutions.

2.2.1 The Pauli Exclusion Principle

The Pauli exclusion principle states that "the wavefunction must be antisymmetric with respect to interchange of any two electrons."[29] Equation 2.8 does not satisfy this requirement as the spin of electrons is not considered. Electron spin can be introduced using the concept of spin orbitals. A spin orbital $f(x)$ can be defined where *x* describes both the spin of the electron and the spatial coordinates (*eq. 2.9*).

$$f(x) = f(r)a(w) \text{ or } f(r)b(w) \qquad 2.9$$

where $a(w)$ and $b(w)$ are orthonormal spin functions describing the electron spin as either up or down[30].

2.2.2 *Slater Determinants*

By writing the many-electron wavefunction as a Slater determinant (*eq. 2.10*) made up of spin orbitals it is possible to include electron spin in the product of one-electron wavefunctions.

$$\Psi(x) = \frac{1}{\sqrt{n!}} \begin{vmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_n(x_n) \end{vmatrix} \qquad 2.10$$

Here n represents the number of electrons and $1/\sqrt{n!}$ is a normalisation constant. Equation 2.10 can be simplified and written as 2.11.

$$\Psi(x) = (n!)^{\frac{1}{2}} \det\left|f_1(1) f_2(2) \ldots . f_n(n)\right| \qquad 2.11$$

The Slater determinant can then be used to evaluate the Hartree-Fock energy via the use of variation theory.

2.2.3 *Variation Theory*[b]

The Slater determinant for an antisymmetric, ground-state, normalised electronic wavefunction can be written, using Dirac notation, as follows (*eq. 2.12*)

$$\left|\Psi_0\right\rangle = \left|f_1 f_2 \ldots f_n\right\rangle \qquad 2.12$$

where Ψ_0 represents the ground-state wavefunction[31].

The variational principle states that "an approximate wave function has an energy which is above or equal to the exact energy."[32] The equality only holds true if the wave function is exact. Thus any approximations will result in an energy that is greater than the ground state energy but tends asymptotically towards the exact ground state energy as the degree of approximation is reduced.

[b] The proof of the variational principle will not be covered here. For a detailed proof the reader should refer to Appendix B of "Introduction to Computational Chemistry" Edn. 1 by F. Jensen.

If we take the n-electron Slater determinant (*eq. 2.11*) then, using Dirac notation, the ground state electronic energy is given as (*eq. 2.13*)

$$E_0 = \langle \Psi_0 | H | \Psi_0 \rangle \qquad 2.13$$

Thus the variational principle can be used to find the determinant Ψ_0 for which the energy E is a minimum. This leads to the formation of the Hartree-Fock equations which allow the energy to be found using an iterative method known as the Self-Consistent Field (SCF) approach.

2.2.4 The Hartree-Fock Equations

The Hartree-Fock equations can be derived using functional variation. The determinant Ψ_0 for which the energy E is a minimum (*eq. 2.13*) is found by seeking a solution such that a small change $\Psi \rightarrow \Psi + d\Psi$ yields no change in the value of E to first order in $d\Psi^2$.

For an infinitesimal change $d\Psi$ the energy is given by (*eq. 2.14*)

$$E[\Psi + d\Psi] = \langle \Psi | H | \Psi \rangle + d \langle \Psi | H | \Psi \rangle \qquad 2.14$$

Since the E is expressed in terms of spinorbitals there exists the constraint that they must be orthogonal and normalised. This constraint is of the form (*eq. 2.15*)

$$g = \sum_{i,j=1}^{n} \left[\langle f_i \mid f_j \rangle - d_{ij} \right] \qquad 2.15$$

Therefore when the spinorbitals are changed by an amount df g changes by (*eq. 2.16*)

$$dg = \sum_{j=1}^{n} d \langle f_i \mid f_j \rangle \qquad 2.16$$

since d_{ij} is constant. It is then possible to take the constraints into account by the use of Lagrange multipliers such that (*eq. 2.17*)

$$dE - \sum_{i,j=1}^{n} l_{ij} d \langle f_i \mid f_j \rangle = 0 \qquad 2.17$$

where l_{ij} is the Lagrange multiplier. By introducing the concept of a coulomb and exchange operator, where *J* is the coulomb operator representing electron repulsion and *K* is the exchange operator representing the exchange of electrons that leads to bonding (*eq. 2.18*)

$$J_u(1)f_a(1) = f_a(1)\int f_u^*(2)\left(\frac{e^2}{4pe_0 r_{12}}\right)f_u(2)\,dx_2$$
$$K_u(1)f_a(1) = f_u(1)\int f_u^*(2)\left(\frac{e^2}{4pe_0 r_{12}}\right)f_a(2)\,dx_2 \qquad 2.18$$

we can form an eigenvalue eigenfunction equation (*eq. 2.19*) that will yield the orbital energies of the spin orbitals.

$$h_1 f_i(1) + \sum_{j=1}^{n}\left[J_j(1)f_i(1) - K_j(1)f_i(1)\right] = \sum_{i=1}^{n} l_{ij}f_i(1) \qquad 2.19$$

An application of the Hartree-Fock procedure for individual spinorbitals can then be formed that yields the energy of each spin orbital (*eq. 2.20*)

$$f_1 f_a(1) = e_a f_a(1) \qquad 2.20$$

where e_a is the spinorbital energy and f_1 is the Fock operator (*eq. 2.21*)

$$f_1 = h_1 + \sum_u\left[J_u(1) - K_u(1)\right] \qquad 2.21$$

Where h_1 is the core Hamiltonian for electron 1, J_u is the coulomb matrix and K_u is the exchange matrix (*eq. 2.22*).

$$J_{ab} = \sum_{cd} D_{cd}\left(f_a f_b \mid f_c f_d\right)$$
$$K_{ab} = \sum_{cd} D_{cd}\left(f_a f_c \mid f_b f_d\right) \qquad 2.22$$

and D_{cd} is the density matrix (*eq. 2.23*)

$$D_{cd} = 2\sum_{m} C_{dm} C_{cm}^{*} \qquad 2.23$$

Each spinorbital can then be obtained from equation 2.20 with the corresponding Fock operator f_i. The fundamental problem, however, is that the Fock operator depends on the spinorbitals of all the other n-1 electrons. Thus it is necessary to solve the equations in an iterative fashion that stops when the solutions become self consistent.

2.2.5 Closed Shell Systems

For a closed shell system, all electrons are paired so it is possible for the HF-SCF procedure to assume that the spatial components of the spinorbitals are identical for each member of an electron pair. This leads to there being, for a system containing n electrons, $\frac{n}{2}$ spatial orbitals of form $\Psi_a(r_1)$ yielding a Restricted Hartree-Fock wavefunction (*eq. 2.24*)

$$\Psi_0 = (n!)^{\frac{-1}{2}} \det\left|\Psi_a^a(1)\Psi_a^b(2)\Psi_b^a(3)\ldots\Psi_z^b(n)\right| \qquad 2.24$$

2.2.6 Open Shell Systems

For an open shell system all the electrons are not necessarily paired in the orbitals such that equation 2.24 no longer holds. There are two common methods for dealing with open shell systems. The first is to use Restricted Open-Shell Hartree-Fock (ROHF). Here all electrons except those occupying open shell orbitals are forced to occupy doubly occupied spatial orbitals. Thus for, e.g. Lithium, the Hartree-Fock wavefunction is (*eq. 2.25*)

$$\Psi_0 = (6)^{\frac{-1}{2}} \det\left|\Psi_{1s}^a(1)\Psi_{1s}^b(2)\Psi_{2s}^a(3)\right| \qquad 2.25$$

There is, however, severe constraints imposed by this formalism. The major problem is that while the 1sα electron has an exchange interaction with the 2sα electron the 1sβ electron does not. A big advantage of this method, however, is that it is an eigenfunction of S^2 hence fundamentally the wavefunction, in terms of the Slater determinant, is correct. Thus anything calculated from this wavefunction will be correct within the constraints imposed.

The second method is the use of Unrestricted Hartree-Fock (UHF). Here the electrons are not implicitly paired and so aren't constrained to the same spatial wave-function. Thus the Hartree-Fock wavefunction for an unrestricted system is (*eq. 2.26*)

$$\Psi_0 = (n!)^{\frac{-1}{2}} \det\left|\Psi_a^a(1)\Psi_b^b(2)\Psi_c^a(3)\ldots\right| \qquad 2.26$$

UHF gives a lower energy than ROHF but has the disadvantage that it is not an exact eigenfunction of S^2 hence it is possible to get spin contamination. This means that any properties calculated from the UHF result that depend on spin (NMR, ESR etc.) will not necessarily be correct.

2.2.7 The Roothaan-Hall Equations

In the early days of quantum chemistry interest was centred, due to the lack of computational power, on solutions of equation 2.20 for atoms. The spherical symmetry of atoms means that the Hartree-Fock equations can be solved numerically to find the spinorbitals. In the case of molecules, however, this symmetry does not exist hence numerical solution of the HF equations is not possible.

A solution to the problem was devised by C.C.J. Roothaan[33] and G.G. Hall[34] who in 1951 independently derived a method for solving the HF equations for a molecule by expanding the molecular orbitals f_a as linear combinations of known basis functions c_j (*eq. 2.27*).

$$f_a = \sum_{j=1}^{N} c_{ji} c_j \qquad 2.27$$

The Hartree-Fock equations (*eq. 2.20*) can then be rewritten as:

$$f_i \sum_{j=1}^{N} c_{ji} c_j = e_i \sum_{j=1}^{N} c_{ji} c_j \qquad 2.28$$

Thus the problem of calculating the wavefunctions has been reduced to computing the coefficients c_{ji}. Multiplication of equation 2.28 by the basis function f_a followed by integration over dr_1 yields the *Roothaan-Hall* equations for a closed shell system (*eq. 2.29*).

$$\sum_{j=1}^{N} c_{ji} \int c_i^*(1)\, f_1 c_j(1)\, dr_1 = e_i \sum_{j=1}^{N} c_{ji} \int c_i^*(1)\, f_1 c_j(1)\, dr_1 \qquad 2.29$$

These are the Fock equations in the atomic orbital basis. It is possible to use a more compact notation by collecting all the N equations from 2.29 and expressing them in matrix form (*eq. 2.30*).

$$Fc = Sce \qquad 2.30$$

where e represents the orbital energies, c the coefficient matrix and S and F the overlap and Fock matrices respectively (*eq. 2.31*).

$$\begin{aligned} S_{ij} &= \langle c_i \mid c_j \rangle \\ F_{ij} &= \langle c_i \mid f \mid c_j \rangle \end{aligned} \qquad 2.31$$

By drawing on the properties of matrix equations it can be shown that the Roothaan-Hall equations only have a non-trivial solution if the secular equation (*eq. 2.32*) is satisfied.

$$\det\left|F - e_i S\right| = 0 \qquad 2.32$$

Unfortunately equation 2.32 cannot be solved directly since the Fock matrix F is composed of coulomb and exchange matrices (*eq. 2.22*), the values of which depend on the spatial wavefunctions that we are trying to find. Thus as before an iterative procedure is required to solve this. This procedure, illustrated graphically in figure 2.1, involves taking a trial set of spinorbitals which are used to construct the Fock matrix. The HF equations are then solved to obtain a new set of spinorbitals which are fed back into the Fock matrix and so on. The cycle is repeated until pre-defined convergence criteria are fulfilled.

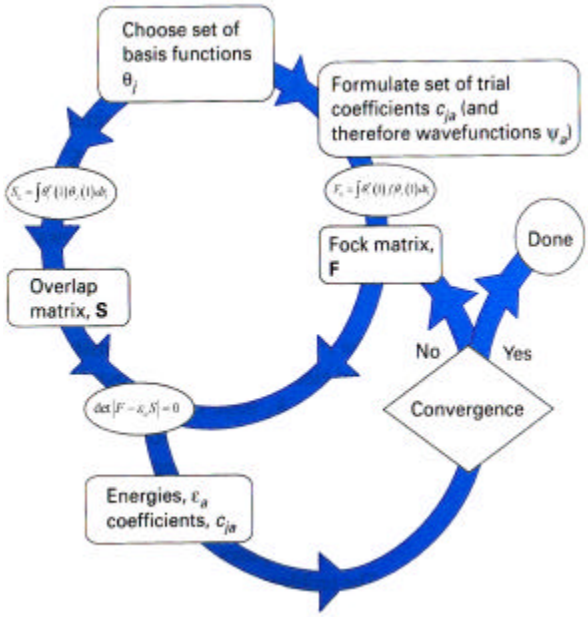


Figure 2.1: Schematic of the iterative procedure in the Hartree-Fock self-consistent field method. Adapted from Atkins, "Molecular Quantum Mechanics" Edn. 3, p. 283.

2.3 Basis Sets

By introducing the Roothaan-Hall equations to solve the HF-SCF procedure it was necessary to express the molecular orbitals f_a as linear combinations of known basis functions c_j (*eq. 2.33*).

$$f_a = \sum_{j=1}^{N} c_{ji} c_j \qquad 2.33$$

In order to form the Overlap and Fock matrices it is necessary to choose a function that represents the form of c_j.

If an infinite number of basis functions were used to represent the molecular orbitals it is clear from the discussion above that the wavefunction solution would essentially be correct, allowing for the neglect of electron correlation. This is known as the Hartree-Fock limit. In practice it is not possible to solve for an infinite number of basis functions so basis sets are chosen to describe the atomic orbitals to varying degrees of accuracy. The choice of basis set is an important factor in electronic structure calculations and fundamentally controls the accuracy of the results that can be expected.

2.3.1 Basis Functions

In choosing the type of basis function used to describe the molecular orbitals there are two practical considerations that have to be taken into account. The first is the accuracy to which the function describes the orbital and the second is the speed with which the two-electron integrals can be evaluated.

2.3.1.1 Slater Functions

The most efficient and accurate functions to use, such that equation 2.33 requires the fewest possible terms for accurate representation of the molecular orbital, are Slater functions[35] which have the functional form, in spherical harmonics, given in equation 2.34.

$$c_{V,n,l,m}(r,q,J) = NY_{l,m}(q,J)r^{n-1}e^{-Vr} \qquad 2.34$$

Where: N is a normalisation constant.
$Y_{l,m}$ is the spherical harmonic function.

Slater functions (*figure 2.2*) have a finite value at zero which correctly describes the cusp at the nucleus. Slater type functions also decay slowly and so represent the wavefunction accurately far from the nucleus. The correct solution to the hydrogen atom is of Slater form, hence using Slater functions is akin to using hydrogen atomic orbitals. However, use of Slater functions leads to difficulties in the evaluation of the one- and two-electron integrals for polyatomic systems. Thus for systems containing four or more atoms the use of Slater functions is very inefficient.

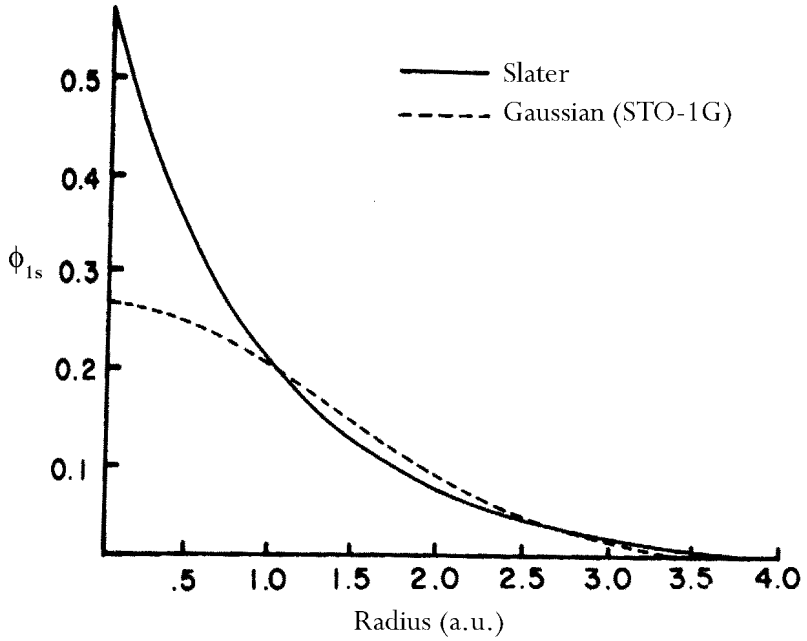


Figure 2.2: Comparison of a Slater function with a Gaussian function - Least squares fit of a 1*s* Slater function by a single STO-1G 1*s* Gaussian function. Adapted from A. Szabo & N.S. Ostlund, "*Modern Quantum Chemistry*", Edn. 1 rev, 1996, p. 157.

2.3.1.2 Gaussian Functions

In 1950 S.F. Boys proposed the use of Gaussian type functions in place of Slater functions to make evaluation of the one- and two-electron integrals more computationally efficient[36].

Gaussian type functions have the functional form, in spherical harmonics, given in equation 2.35.

$$c_{V,n,l,m}(r,q,J) = NY_{l,m}(q,J)r^{(2n-2-l)}e^{-Vr^2} \qquad 2.35$$

Where: N is a normalisation constant.
$Y_{l,m}$ is the spherical harmonic function.

Gaussian type orbitals, as shown in figure 2.2 above, differ from Slater type orbitals in two respects. Firstly at a distance of zero from the nucleus a Gaussian type function has zero gradient while a Slater type function has a finite gradient. Thus Gaussian type functions fail to describe the cusp at the nucleus correctly. Secondly at large distances from the nucleus Gaussian type functions tail off with an exponential quadratic dependence while Slater type functions show a linear dependence. Thus at first sight it would appear that Slater type functions are vastly superior to Gaussian type functions.

However, despite these disadvantages Gaussian type functions have one big advantage over Slater type functions; they allow the one- and two-electron integrals to be evaluated much faster by exploiting the Gaussian product theory. This theory states that the product of two Gaussians on different centres is a single Gaussian situated on a third centre (*fig. 2.3*). This allows the three- and four-centre two-electron repulsion integrals to be reduced to two-centre integrals so giving Gaussian type functions a massive computational advantage over Slater type functions.

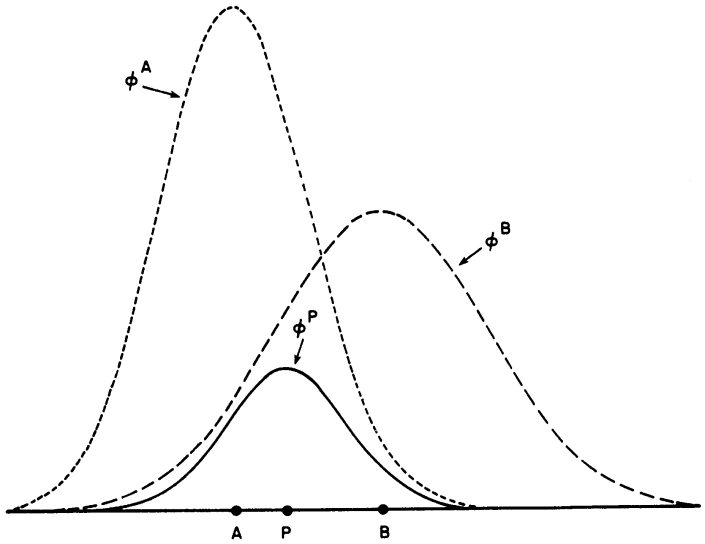


Figure 2.3: Illustration of the Gaussian product Theory - The product of two 1*s* Gaussian functions centred on points A and B is a third 1*s* Gaussian centred on point P. Adapted from A. Szabo & N.S. Ostlund, "*Modern Quantum Chemistry*", Edn. 1 rev, 1996, p. 155.

This leads to the dilemma that the ideal representation would be to use Slater type functions to represent the molecular orbitals, however, Gaussian type orbitals are significantly easier to

evaluate. Fortunately there exists a solution that involves constructing a basis function from a linear combination of Gaussian type functions. Such linear combinations are called contractions and the resulting function is termed a contracted Gaussian function (*eq. 2.36*).

$$c_j^{CGF} = \sum_{p=1}^{L} d_{pj} c_j^{GTF} \qquad 2.36$$

Where: L is the length of the contraction.
d_{pj} is a contraction coefficient.

2.3.2 Standard Basis Sets

Standard basis sets use contracted Gaussian type functions to represent Slater type orbitals. The number of primitive Gaussian type orbitals that are used to form each contracted Gaussian type function reflects on the accuracy of the basis set. The more primitive Gaussians used the more accurate the orbital description is. This is illustrated in figure 2.4 which shows the quality of the least squares fit of a 1*s* Slater function using 1, 2 and 3 primitive Gaussians.

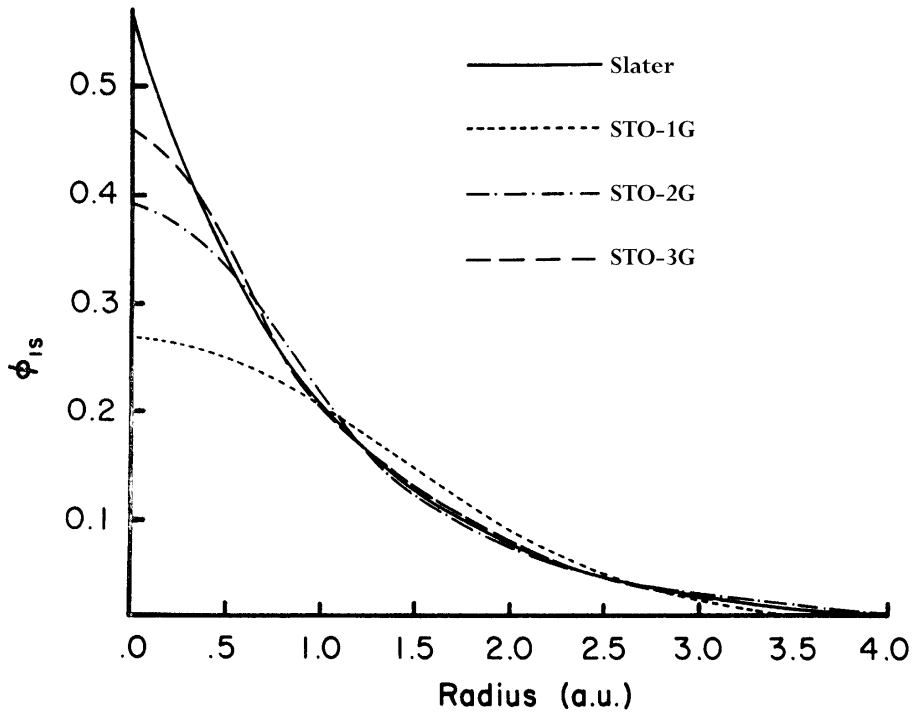


Figure 2.4: Comparison of the quality of the least squares fit of a 1*s* Slater function obtained at the STO-1G, STO-2G, and STO-3G levels. Adapted from A. Szabo & N.S. Ostlund, "*Modern Quantum Chemistry*", Edn. 1 rev, 1996, p. 158.

Typically combinations between one and 6 Gaussian primitives are used to form a single Slater type orbital, termed a basis function. The contraction coefficients used in each case are those that give the best fit to a Slater type orbital.

2.3.2.1 Minimal Basis Sets

A minimal basis set is the least expensive basis set that can be used to describe the molecular orbitals for the system. It is minimal in the sense that it has the least number of functions per atom required to describe the occupied atomic orbitals of that atom.

The general form of a minimal basis set is STO-LG where *L* is the number of Gaussian primitives contracted to form a basis function. An example is the STO-3G basis set by Pople *et al.*[37] which uses 3 primitive Gaussians for each of the inner and valence shell orbitals of each atom. For first row elements and hydrogen, the basis set is denoted by (6*s*3*p*/3*s*) primitives contracted to [2*s*1*p*/1*s*]. Thus for Hydrogen and Helium there is 1 basis function (1*s* orbital), for Li to Ne 5 functions (1*s*, 2*s*, $2p_x$, $2p_y$, $2p_z$), for Na to Ar 9 functions etc.

The unique aspect of the STO-LG basis sets is that in the least squares fit to the Slater type orbitals the contraction exponents are constrained to be identical so that the 2s and 2p fits are performed simultaneously. Thus the 2s and 2p functions have the same radial behaviour and hence can be treated as one function in the integral evaluation step which leads to a very efficient evaluation procedure.

Minimal basis sets are relatively inexpensive and so can be used for calculation on quite large molecules[38]. However, they are so small that at best can only offer qualitative results, but they do include the essentials of chemical bonding.

2.3.2.2 Double Zeta Basis Sets

A significant improvement in results can be obtained by adopting a double-zeta (DZ) basis set instead of a minimal STO-LG basis set. In a DZ basis set each STO is replaced by two STOs that have different contraction coefficients. This gives more flexibility to the size of the orbitals being described and hence gives a more accurate solution. However, this representation doubles the number of basis functions that need to be evaluated and hence increases the number of integrals that need to be solved (analogous to computational complexity), by 2^4.

For the first row atoms the Dunning DZ basis set[39] consists of (9*s*5*p*) primitives contracted to [4*s*2*p*]. It is of course also possible to have higher orders such as triple zeta basis sets that have three functions per orbital etc.

2.3.2.3 Split Valence Basis Sets

Split valence basis sets offer a compromise between the inadequacy of STO-LG minimal basis sets and the high computational demand of double or triple zeta basis sets.

In a split valence basis set such as 3-21G[40] or 6-31G[41] the valence shell atomic orbitals are described by two basis functions, as in the double zeta example above, while each core atomic orbital is represented by a single basis function. This is a reasonable approximation to make as it is generally only the valence shells that are responsible for the chemical properties of an atom or molecule.

In the 3-21G basis set, 3 primitive Gaussians, contracted to a single basis function, are used to describe the core orbitals while 2 primitives, contracted to a single basis function, and an additional, more diffuse, function collectively describe the valence orbitals. For first row atoms this basis set consists of (6*s*2*p*) primitives contracted to [3*s*2*p*] (9 basis functions).

Similarly the 6-31G basis set uses 6 primitives for the core orbitals and 2 functions composed of 3 primitives and a single primitive to describe the valence orbitals. The contraction here, for first row atoms, is (10*s*4*p*) to [3*s*2*p*].

Along the same lines as discussed in 2.3.2.2 it is possible to have triple split valence basis sets such as 6-311G[42] where the valence orbitals are described by three functions, formed from a contraction of 3, 1 and 1 Gaussian primitives respectively, and the core orbitals by one function formed from a contraction of 6 Gaussian primitives. For the first row atoms the contraction is (11*s*5*p*) to [4*s*3*p*].

2.3.2.4 Extended Basis Sets

Up to this point the basis sets that have been described have not allowed for the possibility of contributions from orbitals representing higher angular momentum values such as d or f functions. In order to account for the distortion of electron density due to other adjacent atoms such higher angular momentum functions are essential.

Extended basis sets offer an improvement over split valence basis sets by including additional functions to describe the valence electrons. The extra functions that are added can be in the

form of polarisation or diffuse functions. Polarisation function basis sets such as 3-21G*[43] and 6-31G*[44] work in the same way as split valence basis sets but also add additional polarisation functions. 3-21G* adds *d*-type functions to all first row atoms while 6-31G* adds these functions to all heavy atoms. The 6-31G**[45] basis set also adds *p*-type functions to hydrogen. As discussed above basis sets of this type are required to describe the atom correctly in a non-uniform electric field that arises from the non spherical environment of a molecule.

It is also possible to have diffuse functions included, large *s*- and *p*-type functions, that allow the valence orbitals to occupy a greater region of space. An example of this type of basis set is the 6-31++G[46] basis set. Inclusion of diffuse functions in the basis set has the effect of reducing basis set superposition error (BSSE) in binding energy calculations. This is a phenomenon that results in the stabilisation of one monomer in a donor-acceptor-complex being over estimated due to the presence of basis functions on the other[47]. While this is one way of reducing the error in binding energy calculations it has the disadvantage that the inclusion of diffuse functions increases the calculation complexity considerably. Thus for relatively large systems such as the heme unit of myoglobin their use is not really practical. An alternative method for correcting for BSSE is given in section 2.7 of this report.

2.4 Electron Correlation

2.4.1 Electron Correlation Energy

As mentioned in section 2.3 Hartree-Fock theory does not account for the instantaneous electron - electron repulsions and thus it is possible for electrons to effectively come too close together. This is especially acute in large systems that have a lot of electrons. Calculations carried out with complete or infinite basis sets, while at the Hartree-Fock limit, will therefore be subject to error known as the correlation energy. This energy is defined as the difference between the exact non-relativistic energy of a system E_0 and the energy at the Hartree-Fock limit E_{HF} (*eq. 2.37*).

$$E_{corr} = E_0 - E_{HF} \qquad 2.37$$

It can be shown, using variational theory (*§ 2.2.3*), that E_{corr} is negative since E_{HF} is an upper bound to the exact wavefunction energy. Neglecting electron correlation typically leads to an error of about 1 % of the total Hartree-Fock energy. This is roughly equivalent to the energy of a single covalent bond[48].

2.4.2 Electron Correlation Energy

There exists two types of electron correlation. Dynamical correlation is the instantaneous electron-electron repulsion experienced between two electrons of opposite spin. Non-dynamical correlation is due to degeneracy present in the configuration of electrons in partly filled orbitals[49].

There are corrections for the non-dynamical correlation in Hartree-Fock theory but this does not have a significant effect on the energy of the resulting wavefunction. Improvement of the wavefunction solution is thus only possible by taking into account the dynamical correlation. There exist a number of different methods for including this in a calculation, Møller-Plesset (MP) Perturbation theory accounts for the correlation energy by including higher excitations in the ground state wavefunction as linear corrections[50]. Configuration Interaction (CI) includes electron correlation by noting that a single Slater determinant is insufficient to describe the wavefunction and instead considers the alternative electron configurations that are possible, within a given basis set. Correct weighting of the importance of each of these different configurations then leads to the construction of an accurate, within the basis set approximation, wavefunction.

A third method, and the one that has been utilised in this research, is density functional theory (*§ 2.5*).

2.5 Density Functional Theory

In 1964 Hohenberg and Kohn[51] successfully proved that the ground-state electronic energy wave function is determined completely by the electron probability density $r(x,y,z)$. In other words, there exists a one-to-one mapping between the electron density of a system and its energy. The significance of this lies in the fact that a wave function for an *N*-electron system contains 3 coordinates (4 if spin is included) for each of the electrons. This totals 3*N* coordinates. The electron density is given by the square of the wavefunction, integrated over *N*-1 electron coordinates. This therefore only depends on three coordinates and is independent of the number of electrons. The fundamental point is that, while the complexity of the wavefunction increases rapidly with the number of electrons, the electron density has the same number of variables and hence is independent of the system size.

There exists a problem however. While it can be proven that each different density yields a different ground state energy, the form of the functional connecting the density and the energy is

not known. Ultimately the goal of density functional theory (DFT) methods is to design functionals that connect these two quantities[52].

Kohn and Sham were the first to successfully tackle this problem, leading to Kohn being awarded a 50 % share of the 1998 noble prize for chemistry.

Assuming the Born-Oppenheimer approximation (nuclear - nuclear repulsion is constant) in a method analogous to the Hartree-Fock method it is possible to divide the energy functional into three parts, the kinetic energy, $T[r]$, the potential energy arising from the attraction between nuclei and electrons, $E_{ne}[r]$ and the electron-electron repulsion, $E_{ee}[r]$. The electron-electron repulsion energy can be split further, with reference to Hartree-Fock theory (*§ 2.2.4*), into Coulomb and Exchange parts, $J[r]$ and $K[r]$ respectively. The electronic energy (E_{elec}) can therefore be written as:

$$E_{elec}[r] = T[r] + E_{ne}[r] + E_J[r] + E_K[r] \qquad 2.38$$

$E_{ne}[r]$ and $J[r]$ can be written in terms of their classical expression (*eq. 2.39*)

$$E_{ne}[r] = \sum_a \int \frac{Z_a r(r)}{|R_a - r|} dr$$
$$J[r] = \frac{1}{2} \iint \frac{r(r)\, r(r')}{|r - r'|} dr dr' \qquad 2.39$$

The factor of 1/2 in the expression for *J* allows the integration to be carried out over all space for both variables.

The final term in equation *2.38*, the exchange correlation energy, is a functional of the electron density and accounts for all the non-classical electron-electron interactions. This is the unknown term in equation *2.38*. Its analytical form is unknown hence in order to solve *2.38* approximations must be made to describe $K[r]$.

Initial attempts to find the form of the functionals for the kinetic and exchange energies considered a non-interacting uniform gas. In such a system it can be shown that $T[r]$ and $K[r]$ are given by equation 2.40.[53]

$$T[r] = \frac{3}{10}\left(3p^2\right)^{\frac{2}{3}} \int r^{\frac{5}{3}}(r)\,dr$$
$$K[r] = -\frac{3}{4}\left(\frac{3}{p}\right)^{\frac{1}{3}} \int r^{\frac{4}{3}}(r)\,dr \qquad \textit{2.40}$$

This solution is known as the *Thomas-Fermi-Dirac* (TFD) model.

Unfortunately the assumption of a uniform non-interacting gas is not applicable to atomic or molecular systems. Attempting to apply the TFD model to molecular systems results in an error of 15 to 50 % in the calculated energy. The TFD model is also fundamentally flawed in that it does not predict chemical bonding hence molecules simply don't exist.

The introduction of DFT methods that are applicable to computational chemistry can be attributed to the work of Kohn and Sham[54]. The main problem with TFD models is the poor representation of the kinetic energy. Kohn and Sham devised a way around this problem by splitting the kinetic energy functional into two parts, one that can be found exactly and the other a small correction term[55]. This allows the unknown density matrix to be approximated in terms of a set of single-electrons (orbitals) (*eq. 2.41*).

$$r(r) = \sum_{i=1}^{N} \left| f_i(r) \right|^2 \qquad \textit{2.41}$$

By the use of occupation numbers (ranging from 0 to 1) for each of the orbitals a more accurate form for the kinetic energy can be found. The key to Kohn-Sham theory is that the kinetic energy is calculated under the assumption that the electrons are non-interacting and then corrected. In reality the electrons do interact and the kinetic energy obtained is not exact. However, the difference between the exact kinetic energy and that found by assuming the orbitals are non interacting is relatively small and can be accounted for by absorbing this into an exchange-correlation term.

The Kohn-Sham equations for the one electron orbitals $f_i(r)$ can then be written, in terms of the orbital energies e_i, as (*eq. 2.42*).

$$\left(\frac{\hbar^2}{2m_e}\nabla_1^2 - \sum_{l=1}^{N}\frac{Z_l e^2}{4pe_0 r_{l1}} + \frac{1}{2}\int\frac{r(r_2)e^2}{4pe_0 r_{12}}dr_2 + V_{XC}(r_1)\right) f_i(r_1) = e_i f_i(r_1) \qquad 2.42$$

V_{XC} is the exchange-correlation potential and is a functional derivative of the exchange-correlation energy (*eq. 2.43*).

$$V_{XC}[r] = \frac{dE_{XC}[r]}{dr} \qquad 2.43$$

Thus if the form of E_{XC} is known the exchange-correlation potential can be found by making an initial guess at the density r and then iteratively solving *2.41* until there is no change in the density and exchange-correlation energy, in a fashion analogous to Hartree Fock theory (*§ 2.2.4*).

The results obtained from a Density Functional calculation therefore depend, not only on the basis set chosen, but also on the approximate form of the exchange-correlation term. A number of different functionals exist for representing the exchange and correlation[c]. There exist true DFT functionals such as BLYP, where the B stands for the Becke exchange functional[56] and LYP the Lee, Yang, Parr correlation functional[57].

Since the exchange-correlation is simply a function it is possible, in theory, to use any type of function as long as it accurately describes the system. Thus it is possible to add empirical amounts of Hartree-Fock exchange that have been shown to lead to more accurate results.

Functionals employing this methodology, of which the two functionals used in this research (B3LYP & B3P86) belong, are termed Hybrid DFT functionals.

2.5.1 Computational Performance of Density Functional Theory

In practice DFT calculations involve a computational effort similar to Hartree-Fock. Similarly, like Hartree-Fock, DFT calculations are also one-dimensional: increasing the basis set size allows

[c] The nature by which the approximate functionals are devised will not be covered here. For background information on each of the functionals used the reader is referred to the original published papers.

for a better description of the Kohn-Sham orbitals and hence more accurate results. The power of Density Functional Theory is thus that one can include electron correlation in electronic structure calculations for the same computational cost as Hartree-Fock Theory.

2.5.2 Problems with DFT

The major problem with Density Functional Theory, and one encountered in this research on numerous occasions, is that it is a very empirical, albeit powerful, technique. When the theory fails to work it is unknown why and there is no systematic method for improving it. Functionals that work exceptionally well for one system, e.g. B3LYP with carbon-monoxyheme, can fail to converge at all for a slightly different system, e.g. deoxyheme. All that can really be done is to experiment with different functionals until one is found that works for the system of interest. A universally applicable functional does not currently exist.

2.6 Fast Multipole Methods[d]

Fast Multipole Methods (FMM) are a way to reduce the formal N^4 scaling of HF and DFT calculations to linear scaling in the large system limit. Originally devised for calculating interactions between a system of classical particles interacting via two-body forces[58], an *N*-body problem, FMM has recently been adapted for use with HF and DFT calculations in quantum chemistry.

The fast multipole method works by splitting the problem into near and far fields. The near field is calculated exactly (by direct SCF methods) while the far field is divided into a number of boxes where the interactions between all the charges in one box and all the charges in another are represented by the interaction between two multipoles centred in each box.

As the distance between boxes increases it is possible, for a given level of accuracy, to use larger and larger boxes (*figure 2.5*). Thus for larger systems where the far field is substantial in comparison to the near field the work is reduced from N^2 scaling to something that approaches linear scaling with respect to system size.

[d] Fast Multipole Methods will only be introduced briefly here. For a more thorough discussion the reader is referred to the third year Chemistry MSci literature report "Multipole Methods for Solving Electronic Wavefunctions" by Ross Walker and the references contained therein. An electronic copy of this report is available in Adobe Acrobat[tm] format on the included support CDRom.

The size of the errors introduced into the calculations can be adjusted by varying the size of the boxes and the length at which the multipole expansion representing the interaction between multipoles is truncated.

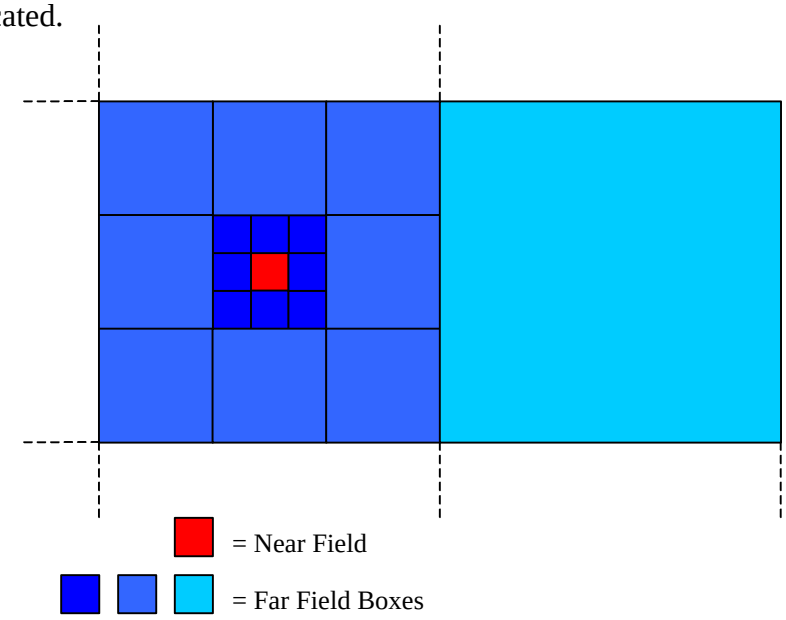


Figure 2.5: Illustration of the hierarchical box structure of the fast multipole method. Adapted from F. Jensen, "Introduction to Computational Chemistry", 1999 p. 387.

2.6.1 FMM Calculation Pre-factors

The mathematics underlying the new fast multipole methods is of 19th century origin hence one is bound to ask why the availability of such methods for sparse decompositions of linear operators has not been recognised until the close of the 20th century. The answer is due to the nature of the scaling of these methods. A method which grows slowly with problem size also, by virtue of the way it works, diminishes less rapidly as the problem size is decreased. Thus there is a minimum problem size, or break even point, at which using a linear scaling (fast) method becomes more economical than traditional slower methods. This is illustrated graphically in figure 2.6. Thus using FMM for calculations on small systems will actually result in the calculation taking longer than standard direct SCF.

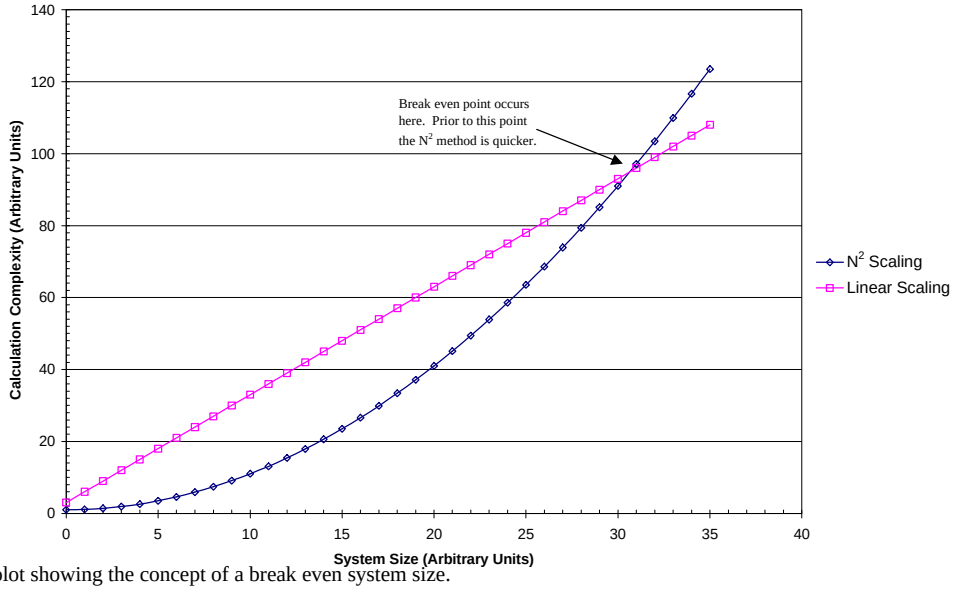


Figure 2.6: Arbitrary plot showing the concept of a break even system size.

In general it is only recently that computers have become powerful enough to study problems of break even size.[59] Thus prior to the 1990's fast multipole methods were purely academic curiosities that were uneconomical for solving problems of the time. It is only recently with the rise in computational power that such methods have become economically viable.

2.6.2 FMM Implementation in Gaussian 98

The computational chemistry program employed in this research, *Gaussian 98 A.7*[60] currently includes FMM routines but only for pure DFT functionals (BLYP, BP86 etc.). The implementation is very recent and has not been tested on a large number of systems. Thus one of the aims of this research has been to investigate the FMM implementation in order to answer the following questions:

1) Does it work and give the correct answer?
2) How fast is it and where is the break even point?
3) How well does it scale
4) Can it be run in parallel?

The results of this investigation are detailed in section *3.1* of this report.

2.7 Binding Energy Calculations

This research has been centred upon the binding of CO and O_2 ligands to the iron atom of the heme unit in myoglobin. The focus has been upon both the predicted structures for each of the bound ligands as a function of the active site structure and the binding energies of carbon monoxide and oxygen to the various deoxyheme structures. The method used for calculating the binding energies is as follows.

2.7.1 Uncorrected Binding Energies

Consider the bimolecular reaction:

$$\mathbf{A + B \rightarrow AB}$$

The geometries of the two isolated molecules can be found yielding an energy $E(A)_a$ for molecule A with the basis functions of a and $E(B)_b$ for molecule B with the basis functions of b. The geometry of the complex can then be found yielding an energy $E(AB)^*_{ab}$ representing the energy of AB with the basis functions of a and b. The geometry of the A and B molecules in the complex will generally differ slightly from the isolated molecules and therefore the complex

geometry is denoted by a *. The complexation (or binding energy) is then given by the dimer energy minus the monomer energies (*eq. 2.44*).

$$\Delta E_{complexation} = E\left(AB\right)_{ab}^{*} - E\left(A\right)_{a} - E\left(B\right)_{b} \qquad 2.44$$

2.7.2 Basis Set Superposition Errors

Unfortunately the simple procedure detailed in section 2.7.1 does not give an accurate estimate of the binding energy due to an effect known as Basis Set Superposition Error (BSSE). For example if we consider the strength of a hydrogen bond between two water molecules the simplest method (assuming a size consistent method is used, e.g. HF, DFT, MP2, full CI) for calculating the binding energy (analogous to bond strength) would be to calculate the energy of the dimer and then subtract twice the energy of the monomer. However, while the electron distribution in each water molecule in the dimer is similar to the monomer the basis functions of the other water molecule in the dimer provide more flexibility for the wavefunction and therefore, assuming the method is variational, will give a lower energy than would be expected. This results in the binding energy being overestimated.

In the infinite basis set limit the BSSE would be zero, however, this is not computationally feasible so a correction must be made to the binding energies to account for the BSSE. A common method of finding the approximate BSSE, and the method employed in this research, is to use the Counterpoise (CP) correction[61].

2.7.3 The Counterpoise Correction

In this method the effect of the additional basis functions are taken into account via the use of four single point calculations that yield a correction to the uncorrected binding energy *(§ 2.7.1*).

The counterpoise correction is found by carrying out 4 additional calculations. The energies of *A* with basis functions *a* and *B* with the basis functions *b* are calculated with the geometry they have in the complex yielding energies $E\left(A\right)_{a}^{*}$ and $E\left(B\right)_{b}^{*}$. Two additional calculations are then carried out on *A* and *B* with the structure they have in the complex but with the inclusion of both basis sets in each case. Thus the energy of *A* with the basis functions *a* and the basis functions of *B* located at the corresponding nuclear positions but without the *B* nuclei present is found and vice versa to yield energies $E\left(A\right)_{ab}^{*}$ and $E\left(B\right)_{ab}^{*}$. The CP correction is then given as (*eq. 2.45*).

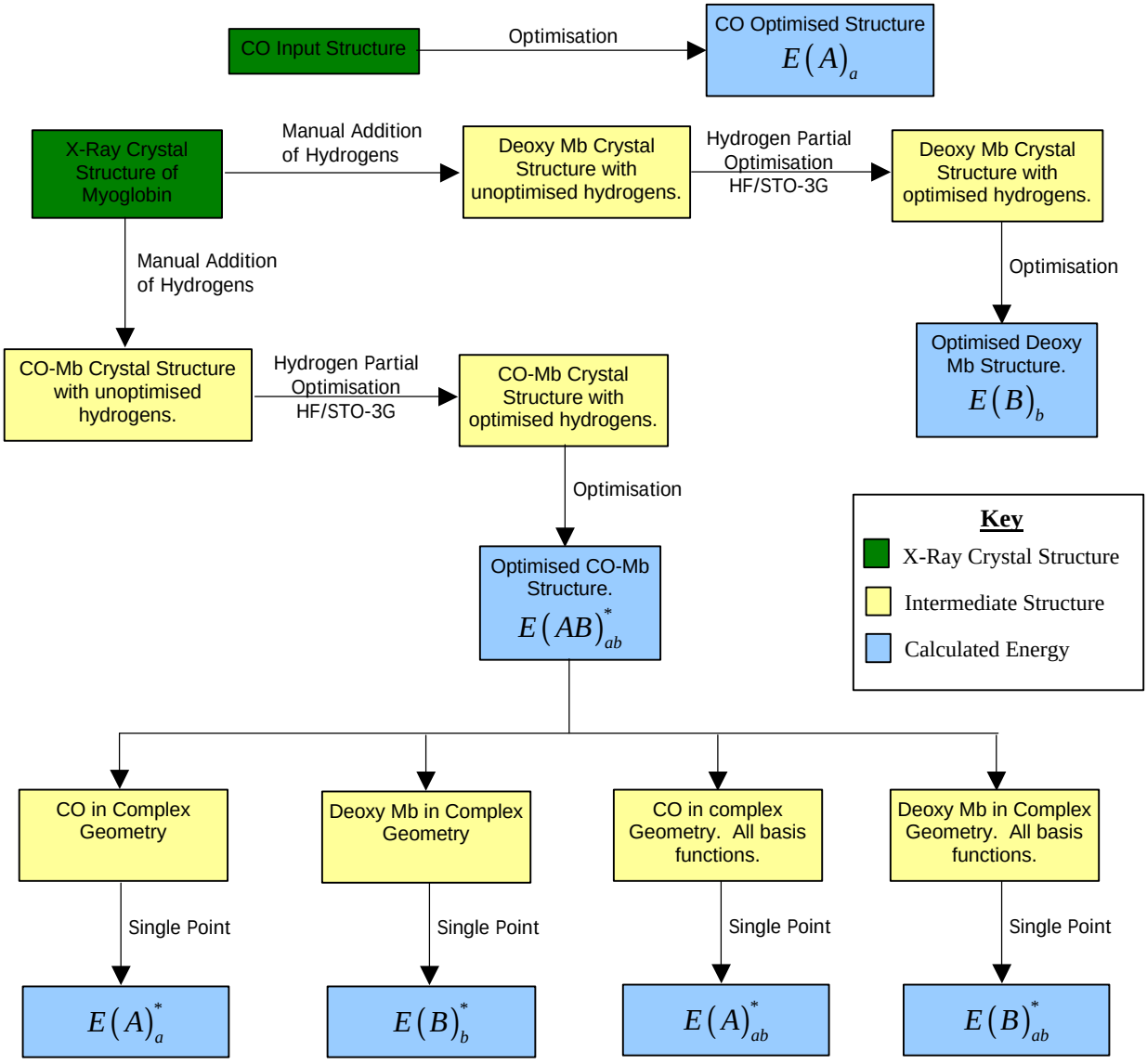
$$\Delta E_{CP} = E(A)^{*}_{ab} + E(B)^{*}_{ab} - E(A)^{*}_{a} - E(B)^{*}_{b} \qquad 2.45$$

The counterpoise corrected binding energy is then given as (*eq. 2.46*).

$$\Delta E_{CP\ Corrected} = \Delta E_{complexation} - \Delta E_{CP} \qquad 2.46$$

The flowchart on the following page (*figure 2.*7) illustrates all the calculations that must be performed to yield a counterpoise corrected binding energy for CO binding to myoglobin. This represents the procedure used for the calculation of binding energies reported in section 5 of this report. The initial input structures are obtained from x-ray crystal structures. The system is then protonated at standard distances and the hydrogen positions optimised using HF/STO-3G while keeping the rest of the system fixed. The resulting structures are then used for the binding energy calculations.

The boxes representing final energies are coloured pale blue while the x-ray data boxes are coloured green and the intermediate boxes are coloured pale yellow.



$$\Delta E_{binding} = E(AB)^*_{ab} - E(A)_a - E(B)_b - E(A)^*_{ab} - E(B)^*_{ab} + E(A)^*_a + E(B)^*_b$$

Figure 2.7: Flowchart illustrating the procedure for calculating the binding energy between carbon monoxide and myoglobin. This procedure is that which has been employed throughout this research and is completely general and can be applied to any ligand-complex system.

3. Initial Scaling & Feasibility Study

Upon starting this research it was unknown how well the available methods would scale with system size and how long the calculations would take on the available resources. This information was essential in deciding how to initially tackle the problem of carrying out a theoretical study of myoglobin. Ideally the best situation would be a complete *ab initio* study using an electron correlation method such as density functional theory with a reasonably large basis set (e.g. 6-31G*). In order to attempt such an ambitious set of calculations a true linear scaling, highly parallel code would be required.

3.1 FMM Scaling Tests

The computational chemistry software *Gaussian 98* used for this research includes a switch for using FMM in pure density functional (non-hybrid functionals) calculations that, in theory, should greatly decrease the computation time required for the study of large molecules. In the words of the *Gaussian* manual[62]:

> "*The cost of computations can be linearized using fast multipole method (FMM) and sparse matrix techniques for certain kinds of calculations.*"

This would suggest that calculations using FMM should show linear scaling in the large system limit. In order to fully study myoglobin using *ab initio* techniques it was essential that the FMM scaled much more favourably than standard DFT methods and that it could run sufficiently well in parallel to make the calculations possible within an acceptable time frame.

At the time of starting this work, however, the FMM routines had not been tried in the department and no published material referring to the performance of the *Gaussian* implementation of the FMM could be found. An investigation was therefore carried out to answer the following questions:

1) Does it work and give the correct answer?
2) How fast is it (where is the break even point)?
3) How well does it scale with system size (is it linear)?
4) Can it be run in parallel?

3.1.1 Procedure

To answer the above questions a number of geometry optimisations were carried out on simple straight chain alkanes running from methane (17 basis functions) to n-pentacontane ($C_{50}H_{102}$, 654 basis functions) using the 6-31G basis set[63] with the Becke exchange functional[64] and the Lee, Yang, Parr correlation functional[65]. Calculations were run both with and without the FMM switch[66] in order to compare performance.

Calculations were initially run sequentially on a single processor and then the largest alkane ($C_{50}H_{104}$) was run in parallel using 2, 3, 4, 5, 6, 7 and 8 processors.

Finally a test job, single point energy minimisation using BLYP/FMM-3-21G, was run on the protein Crambin (642 atoms, 3597 basis functions), approximately a quarter the size of myoglobin, to test the feasibility and resource requirements of a full *ab initio* study.

During the process of these investigations a separate comparison was made between different machine architectures using the same alkane series as detailed above. The details and results of this investigation are reported separately in appendix A at the end of this report.

3.1.2 Computation

The computational chemistry package *Gaussian 98 revision A.7* was used for this investigation. All calculations were run on a shared memory 44 processor MIPS R10000 (195 MHz IP27 4 mb L2 Cache), Silicon Graphics ONYX 2 based in the Centre for Computing Services at Imperial College London. The machine consists of 22 dual processor node boards with 512 MB of local memory each connected via a single SGI Cray Link forming a shared memory pool of 11 GB. The operating system used was *IRIX version 6.5* and the *Gaussian 98* source code was compiled using the *MIPS Pro F77 compiler v7.2*.

All alkane geometries were drawn in *CambridgeSoft Chem 3D v4.0*, initially optimised using the Allinger MM2 molecular mechanics force field[67] and then exported as Cartesian coordinates. DFT Calculations were then run in *Gaussian 98* with the use of symmetry explicitly turned off using the NoSymm keyword. Each job was given a maximum memory allocation of 40 MW (320 MB) and the convergence criteria was left at the default of 10^{-8}. The reported CPU time for each job was then extracted from the output files, converted to seconds and exported as a *Microsoft*

Excel 2000 spreadsheet using custom written software, described in appendix B at the end of this report.

For the crambin test the cartesian input geometry was extracted from the *Gaussian 98* AMBER test job (#448). A single point energy minimisation using BLYP/FMM-3-21G was selected. The job was allocated 128 MW (1024 MB) of memory and run in parallel on 4 processors.

3.1.3 Results and Discussion[e]

In all the alkane calculations the predicted electronic energies were recorded and found to be, for a given alkane using BLYP or BLYP/FMM, the same to within $\pm 10^{-5}$ %. This is a very encouraging result as it means the fast multipole methods can be expected to yield the same results as normal DFT methods.

Plotting the optimisation time, for single processor jobs, against the number of basis functions (*figure 3.1*) shows that the fast multipole method becomes faster than the pure DFT method for as few as 130 basis functions and by 654 basis functions is 1.86 times faster.

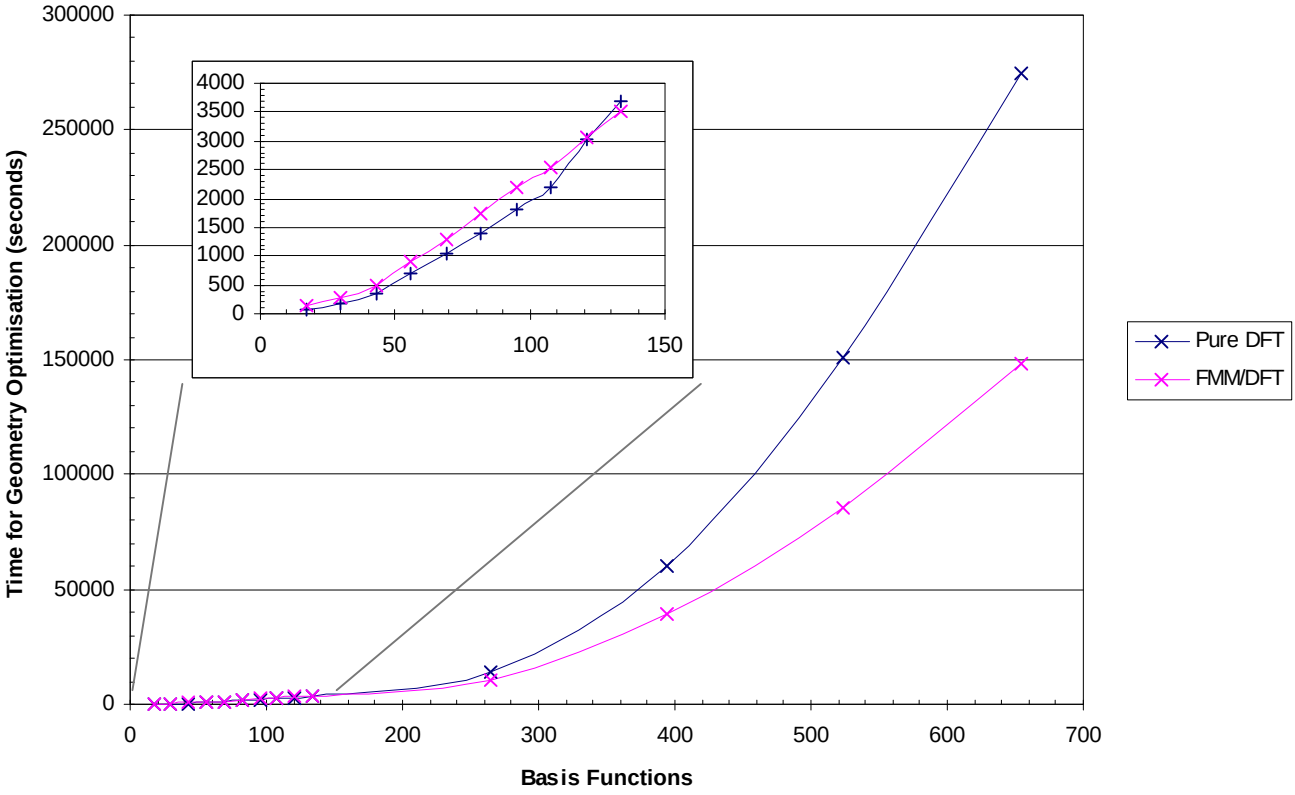


Figure 3.1: Plot of computation time against number of basis functions for geometry optimisation of a range of straight chain alkanes using BLYP and BLYP/FMM methods (6-31G).

[e] In the process of keeping things concise the tabulated results are not reproduced in this report. The numerical results and raw *Gaussian 98* output files are available on the included support CDRom.

Contrary to what is written in the *Gaussian* manual, however, it can be seen that while offering better scaling than pure DFT methods the scaling is still far from linear. It is possible that the scaling may indeed become linear as the system size increases but the resources were not available to investigate this fully.

The results obtained for the multiprocessor jobs on *n*-pentacontane are shown below (*figure 3.2*).

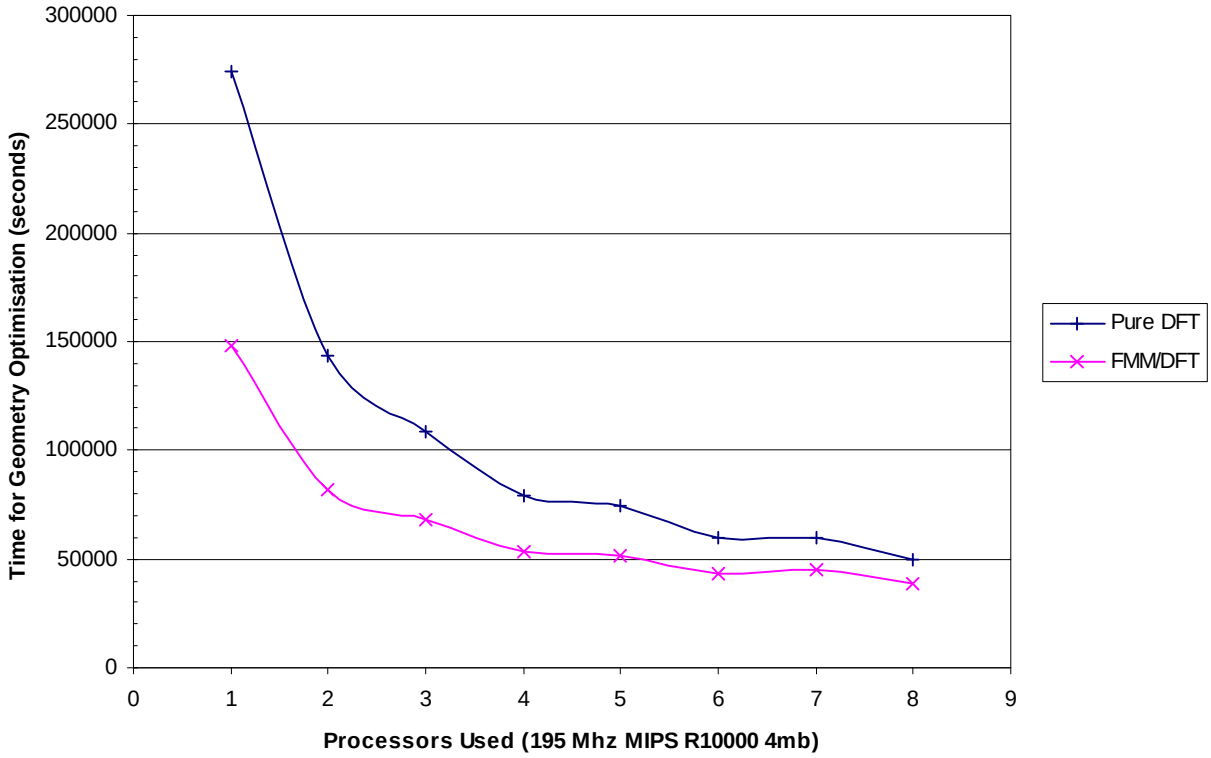


Figure 3.2: Plot of computation time against number of processors used for geometry optimisations of *n*-pentacontane (6-31G) using BLYP and BLYP/FMM.

From figure 3.2 it can be seen that the time taken is shortened by using multiple processors but the effect becomes asymptotic at approximately 4 processors. The reason for this is due to the increase in communication overhead on going to more and more processors. The fact that the speed increase becomes asymptotic at 4 processors is due in part to the machine architecture but also to the way in which parallel processing is implemented in *Gaussian 98*. It is also interesting to note that the bumps present in both curves in figure *3.2* correspond to odd numbers of processors. It would appear that the *Gaussian 98* SMP routines perform better with even numbers of processors. Exactly why this is the case is unknown at the current time but may be worth investigating in another project. What is not obvious from figure 3.2 is that the FMM routines do not scale to multiple processors as well as the standard DFT routines. By normalising the results using equation 3.1 it is possible to plot a graph showing the efficiency of the FMM routines against the DFT routines as a function of the number of processors used (*figure 3.3*).

$$\text{Normalised Time} = \frac{\text{Time for FMM Calculation}}{\text{Time for Pure DFT Calculation}} \qquad 3.1$$

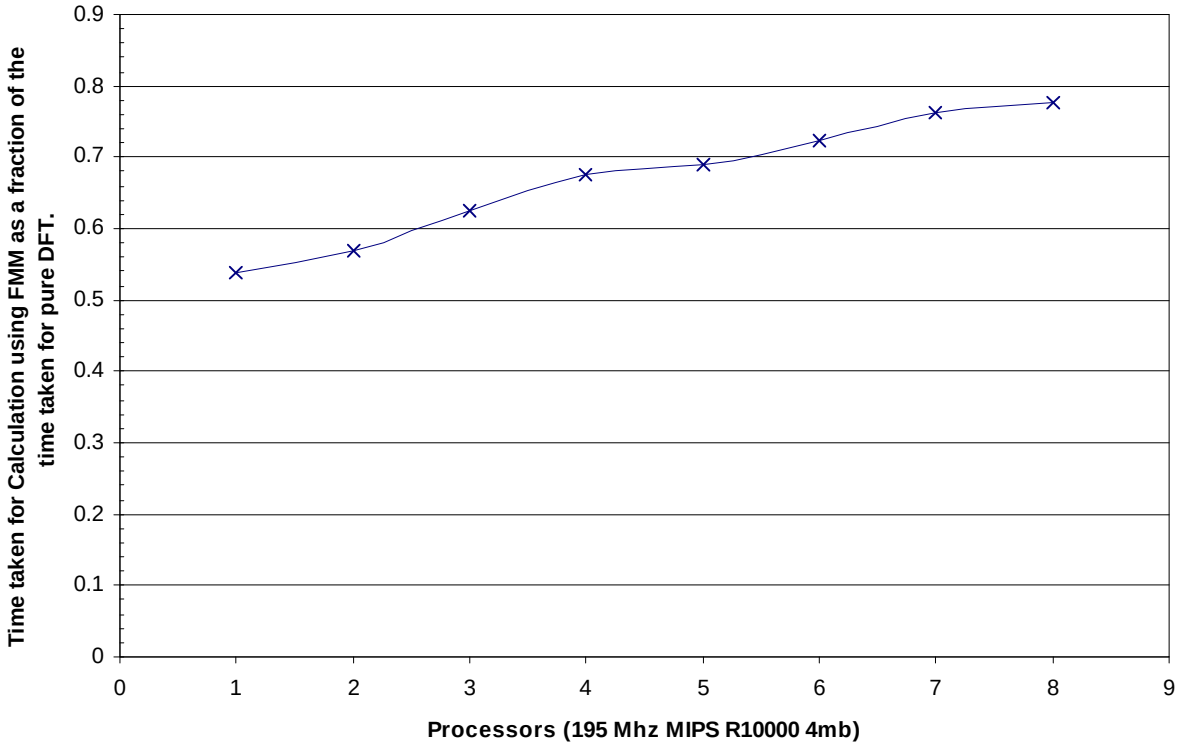


Figure 3.3: Plot showing FMM time / DFT time as a function of the number of processors used for geometry optimisation of *n*-pentacontane (6-31G).

From figure 3.3 it can be seen that the standard DFT routines in *Gaussian 98* scale to multiple processors more efficiently than the FMM routines and the trend continues as the number of processors is increased. Ultimately there will come a point where, due to the poor parallel implementation of the FMM, the standard DFT methods will become more efficient.

In conclusion these results, while showing that the FMM implementation in *Gaussian 98* does indeed perform more efficiently than the standard DFT methods, indicate that the scaling is still too steep and the scalability to multiple processors too poor to make a full *ab initio* study of myoglobin feasible within the resource limits and time frame of this research. This conclusion is further verified by the crambin test calculation results given in section 3.1.3.1 below.

3.1.3.1 Crambin Test Calculation Results

The test calculation on crambin was halted after failing to converge within 64 SCF cycles. In order to reach 64 cycles, however, the calculation took a total of 16 days on 4 processors and required just under a gigabyte of ram. At 3-21G a full calculation on myoglobin would total 14660 basis functions which even optimistically assuming linear scaling with system size

would mean approximately a day per SCF cycle. A full optimisation is therefore well beyond reach with the available resources and software.

The crambin test also highlighted a second problem. The calculation required approximately 960 MB of ram which is just below the value of SHMMAX 0x40000000 (1024 MB) which represents the maximum size of a shared memory segment in *IRIX 6.5*. This limit means that a calculation on myoglobin, which tests show would require at least 3.5 GB, could not be run in parallel without modification of the machines' kernel. Such large memory requirements would also place a huge strain on the Cray memory link between the machines' nodes hence such a large job is likely to be I/O bound rather than processor bound so the improvement in running the job in parallel is likely to be a lot less than that observed for the alkane series above (*§ 3.1.3*).

In conclusion therefore studies of very large systems (> 2000 atoms) using *ab initio* methods are not feasible using current technology unless highly efficient parallel algorithms can be produced for the iterative solution of the SCF calculations.

4. Development of Working Methodology for Study

The aim of this research was to investigate how myoglobin selectively discriminates between carbon monoxide and oxygen. It was therefore decided to look at how the predicted binding energies for CO and O_2 changed with the environment around the heme. The question therefore is whether it is based simply on the two histidine residues close to the iron atom or whether the entire protein plays a part. The investigation in section 3 of this report indicated that a full *ab initio* study is beyond reach with current technology. The decision was therefore made to initially look at just the heme unit and surrounding proximal and distal histidines using an *ab initio* approach.

In order to conduct such a study it was necessary to develop a methodology for studying the system. Initial trial calculations on deoxy and carbon monoxyheme showed some serious short comings with the default Gaussian options for optimisations. The major problem encountered was the poor convergence of all optimisations leading to calculations involving as few as 200 basis functions either not converging at all or taking over 3 weeks to complete on a 4 processor SGI ONYX 2. Thus a detailed investigation of the factors affecting the quality of results and the convergence of the SCF routines was conducted. The options investigated, the results obtained and the overall conclusions are summarised in this chapter.

4.1 Level of Theory

The first step in developing a working methodology was to decide on a level of theory to use. Based on the available computer power there was really only two choices available. Either Hartree Fock or Density Functional Theory. Use of Møller Plesset 2nd Order Perturbation Theory or Configuration Interaction would have proven too computationally intensive to carry out a sufficiently in-depth study within the timescale of this research project.

4.1.1 HF vs DFT

Hartree Fock is the simplest level of theory that can be employed in an *ab initio* calculation so it was decided to start with this and see how well it performed.

Calculations were set up for bare deoxyheme and for deoxyheme and his 93 (proximal). The initial structures were taken from the crystallographic data for deoxy horse heart myoglobin provided by the Brookhaven protein databank (1WLA)[68]. The desired units were then extracted from the PDB file, protonated manually at standard values for the bond lengths and angles[69] and then the hydrogens optimised using HF/STO-3G while keeping the rest of the structure fixed.

4.1.1.1 HF Computation

The structures of deoxyheme and deoxyheme+his93 were optimised at HF/3-21G using *Gaussian 98 Rev. A.7* on a Silicon Graphics ONYX 2 running IRIX version 6.5. To simplify the calculation the histidine (his 93) was truncated at the β-carbon. In order to obtain convergence at the default settings the maximum number of SCF Cycles was increased from 64 to 256 and the convergence limit was reduced from 10^{-8} to 10^{-6}. In both cases the system charge was specified as –2 and the spin state as a singlet.

4.1.1.2 HF Results and Discussion

Using HF with a convergence limit of 10^{-6} the calculations completed after approximately 5 days with few convergence problems. However, on examining the structures obtained it can be seen that, due to the neglect of electron correlation, Hartree Fock performs very poorly especially in the case of heme+his93 where it predicts a structure that is clearly wrong. Figures 4.1 and 4.2 show comparisons between the input structures and the output structures obtained. Full 3 dimensional versions of these structures are available on the support CDRom.

For the bare heme the output structure shows the heme unit with a twist of approximately 5^{o} when really it should be flat. The heme+his 93 structure is even worse showing very bad distortion of the heme unit.

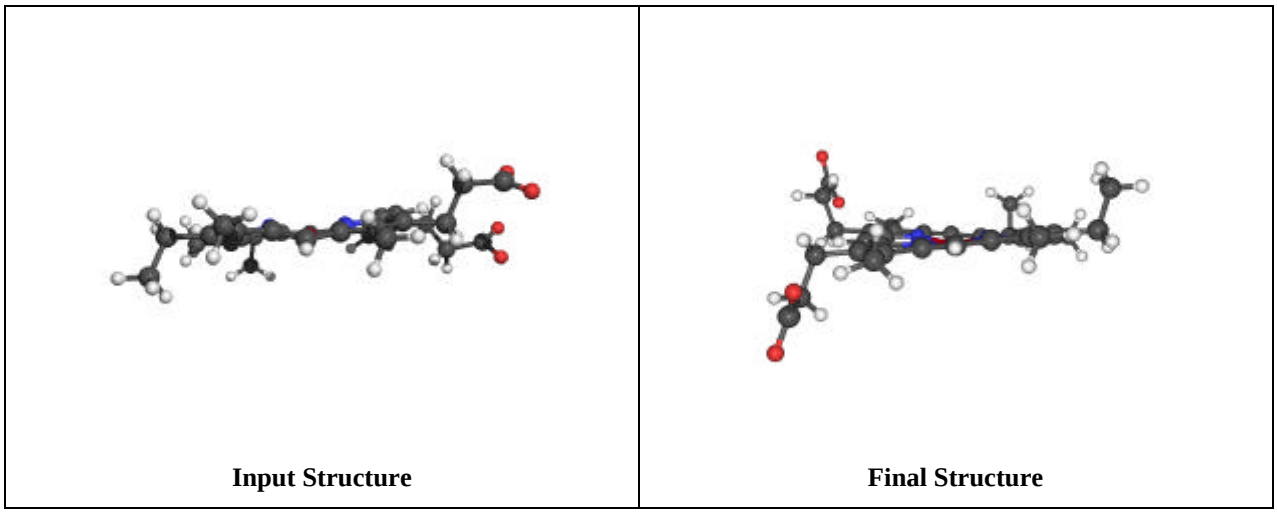


Figure 4.1: Diagram showing input structure and final structure for HF/3-21G optimisation of deoxyheme.

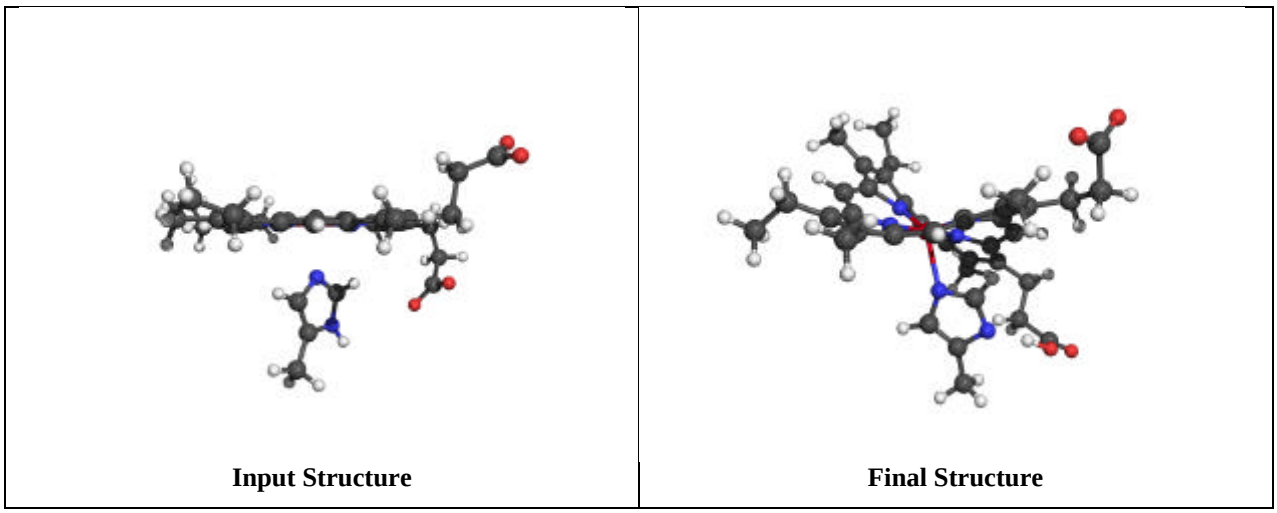


Figure 4.2: Diagram showing input structure and final structure for HF/3-21G optimisation of deoxyheme+his93.

From these results it was concluded that Hartree Fock theory is insufficient to reliably model the behaviour of the heme unit and surrounding residues of myoglobin. The main reason for the poor performance is probably due to the neglect of dynamic correlation. It was therefore decided to try density functional theory instead as this includes electron correlation for approximately the same computational cost as Hartree Fock.

4.1.1.3 DFT Computation

Optimisations of deoxyheme and deoxyheme+his 93 were carried out as detailed in section 4.1.1.1 but using DFT (B3LYP[70, 71]) instead of HF.

4.1.1.4 DFT Results and Discussion

The optimised structures obtained with density functional theory are far better, and closer to those expected, than the Hartree Fock structures. The results obtained are illustrated in figures 4.3 and 4.4 below.

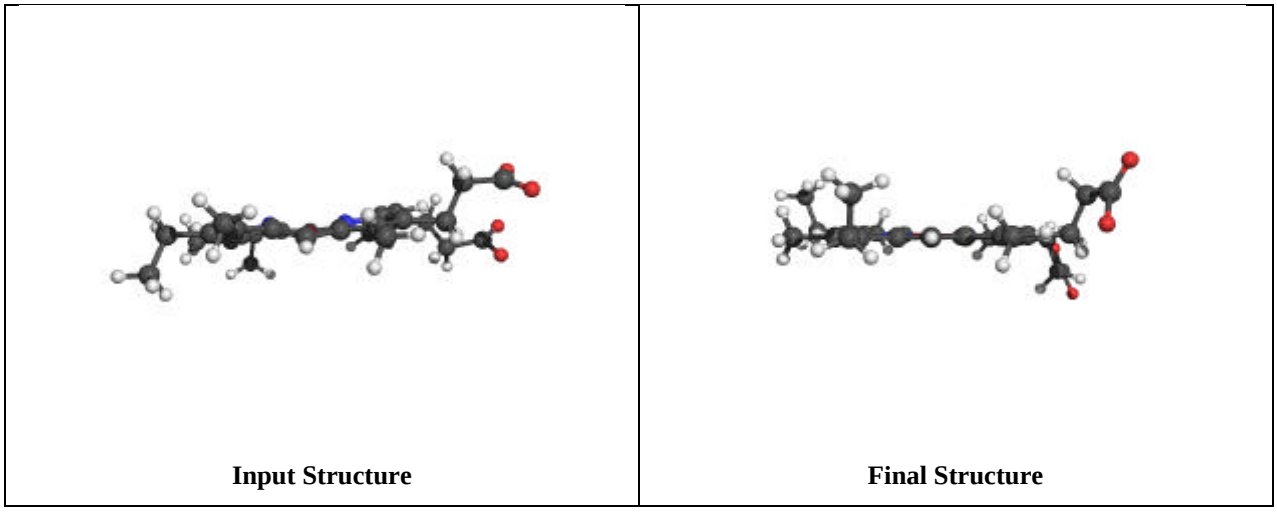
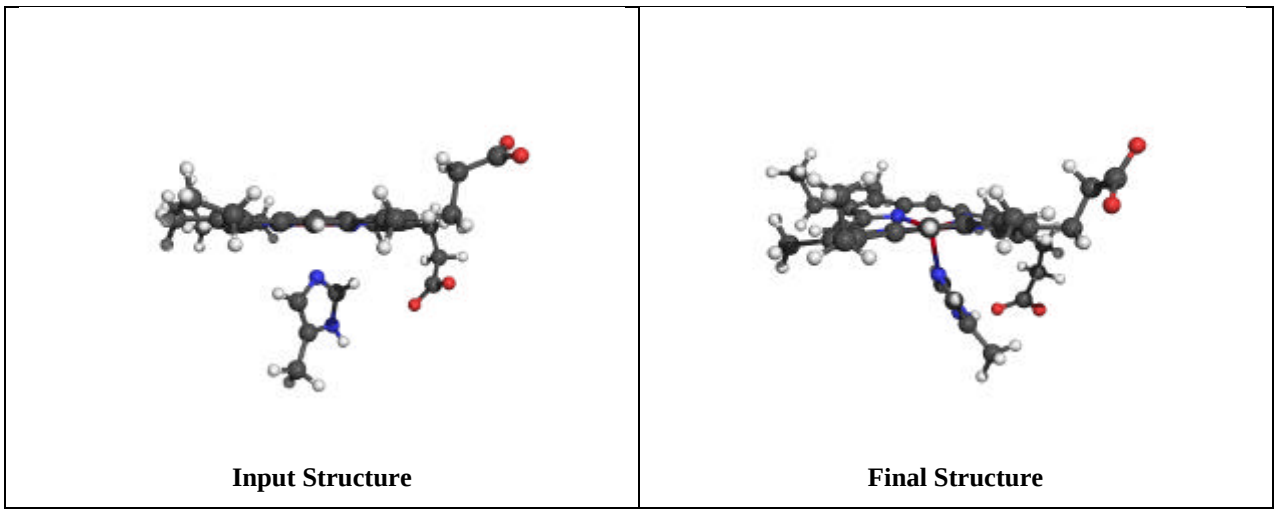


Figure 4.3: Diagram showing input structure and final structure for B3LYP/3-21G optimisation of deoxyheme.



Input Structure **Final Structure**

Figure 4.4: Diagram showing input structure and final structure for B3LYP/3-21G optimisation of deoxyheme+his93.

It is obvious from the above examples that DFT performance is far superior to HF. For the bare heme unit it predicts a flat ring as expected. For the heme+his93 structure it still performs badly predicting a twisted ring, however, the distortion is far less than that obtained with HF. It was found later *(§ 4.8)* that using a tighter convergence of 10^{-8} coupled with the use of larger basis sets, particularly on the iron and pyrrole nitrogens reduces the distortion considerably and leads to more accurate results. From these calculations and other results, not reported here, it was decided to use DFT for all further calculations.

A big problem, however, was found with the use of density functional methods; the convergence problems were severely amplified with DFT with both calculations requiring a total of approximately 4 times more SCF Cycles than the corresponding HF calculations and consequently requiring 4 times more computer time. With a more desirable convergence limit of 10^{-8} convergence is still not achieved after 512 cycles with the Gaussian default options. Use of larger basis sets including 3-21G*, 6-31G* and 6-31G* + Ahlrichs pVDZ[72] for Iron failed to improve the convergence but instead just led to the calculations taking much longer. Thus in order to look at the ligand binding energies with full counterpoise correction it was necessary to look closely at all the available options in order to arrive at a methodology that gave reliable convergence of each optimisation step within 256 cycles. The procedures used and the parameters investigated are discussed in the following sub sections.

4.2 Selection of DFT Functional

The first stage in the development of a working methodology for study was to investigate the effect various DFT functionals had on the number of SCF cycles required for convergence of the calculations. Recent work within the research group[73] has shown that B3LYP often performs poorly for iron based systems. Work by Griffiths suggested that B3P86[74] generally performs

better for transition metal systems. A possible explanation for this is that the LYP correlation functional was parameterised for He, only 4 electrons, zero spin unpaired correlation energy, whereas the P86 correlation functional was parameterised for Ne, more electrons. Perdew estimates that there is 21% spin unpaired correlation present in Ne[75]. Thus a deficiency of B3LYP is that it cannot adequately treat paramagnetic systems such as Fe(II) can be.

4.2.1 Computation

In order to test the effect of the DFT functional on the convergence of the heme calculations and to see if B3P86 was in fact better than B3LYP a set of calculations were set up for CO-Heme. Carbon monoxyheme was used because it was found that convergence could be more reliably achieved if a ligand was present on the iron. The protonated heme unit from section 4.1 was used in these calculations with the CO ligand manually built in at the crystallographic distances suggested by neutron diffraction of sperm whale myoglobin (2MB5[76]).

Single point energy minimisations were performed using *Gaussian 98* on a Silicon Graphics ONYX 2 running irix 6.5. All calculations used the 3-21G* basis set with the maximum number of SCF cycles increased from 64 to 256 and the convergence limit reduced from 10^{-8} to 10^{-6}. The Vshift value (separation of occupied and virtual orbitals) was set at 500 millihartrees while the number of simulated annealing steps (IOP 5/71) was increased from 10 steps to 20 steps as previous test calculations indicated that these settings aided convergence. In all cases the charge was set to –2 and the spin state to a singlet.

The functionals tested consisted of a mixture of pure DFT functionals (BLYP, BP86, BPW91[77-81]) and hybrid functionals (B3LYP, B3P86, B3PW91, G96LYP[82,83]). Each calculation was run in turn on a single processor with the only option changed between calculations being the functional used. On completion of each calculation both the CPU time and the number of SCF cycles required for convergence were recorded.

4.2.2 Results and Discussion

The results obtained are presented in the following two graphs, figure 4.5 shows the number of SCF cycles required for a single point energy minimisation while figure 4.6 shows the CPU time required for each energy minimisation. A tabulated version of these results and the raw Gaussian output files are available on the support CDRom.

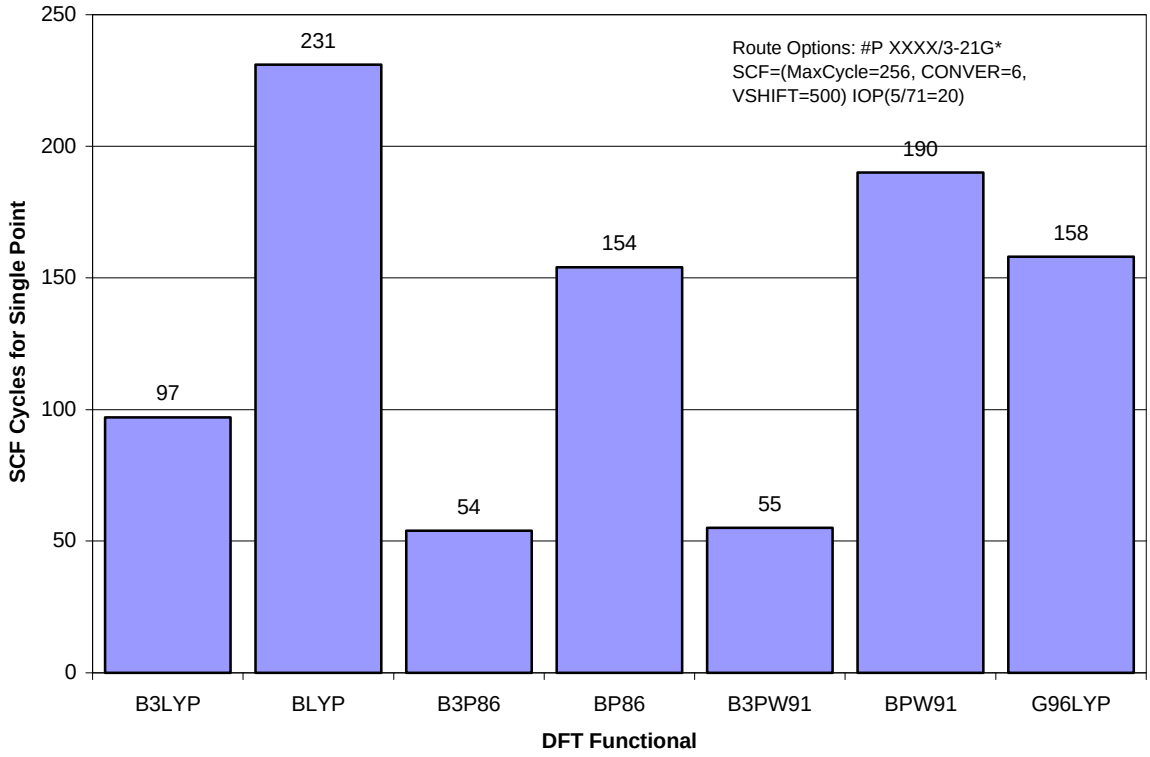


Figure 4.5: Plot showing the number of SCF Cycles required for a single point energy minimisation of Heme-CO for various DFT functionals.

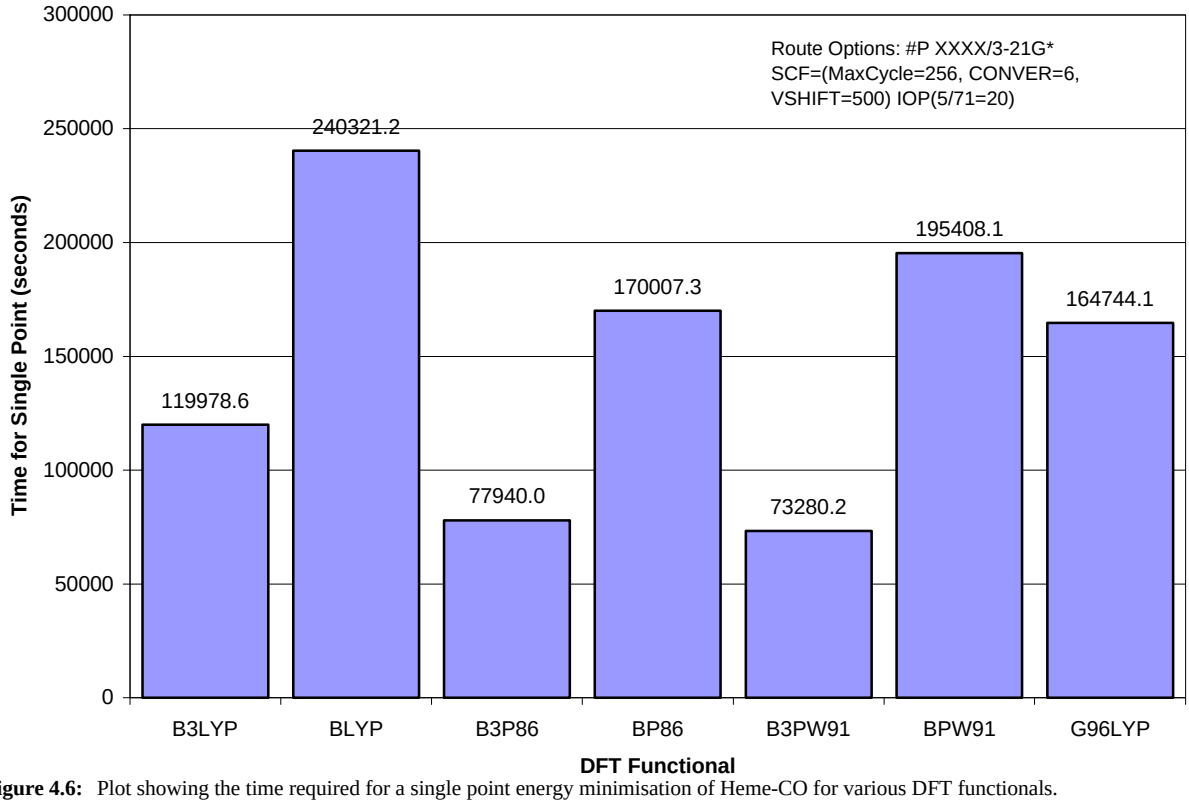


Figure 4.6: Plot showing the time required for a single point energy minimisation of Heme-CO for various DFT functionals.

From the above graph it can be seen that B3P86 does indeed perform better than B3LYP requiring the least number of cycles for convergence, approximately 56 % of the number required for B3LYP. This was closely followed by B3PW91 which required one more cycle than B3P86 but actually completed in only 94% of the time required for B3P86. All the pure DFT functionals

performed very poorly requiring between 3 and 4 times more SCF cycles for convergence than B3P86.

Based on these results B3PW91 would appear to be marginally better, in terms of convergence speed, than B3P86, however, personal correspondence with a number of people in the field suggested that the accuracy of results obtained using B3PW91 had not been sufficiently verified to rely upon whereas B3P86 had. It was therefore decided to use B3P86 for the binding energy calculations.

4.3 Optimisation of Level Shifting

The technique of level shifting[84] is best understood by considering the molecular orbitals which form the basis for the Fock operator (*§ 2.2.4*). At convergence the elements of the Fock matrix in the molecular orbital basis between the occupied and virtual orbitals are zero. The SCF procedure involves taking linear combinations (mixing) of the occupied and virtual orbitals. During the iterative procedure such linear combinations can be large which can often lead to the total energy either increasing or oscillating. The amount of mixing can be reduced by artificially increasing the energy gap between the occupied and virtual orbitals (i.e increasing the virtual orbital energy). It can be shown[85] that if the energy gap is made large enough the total energy is guaranteed to decrease, so forcing convergence. However, while increasing the energy gap makes the convergence more stable it decreases the rate with which convergence occurs. There is therefore a trade off to be made between the amount of level shifting used and the time available for running the optimisation. Too little level shifting and the system can oscillate and never converge, too much and convergence can take a very long time.

With this in mind it was decided to experiment with the amount of level shifting used in an attempt to optimise the convergence process and possibly achieve convergence at the more desirable tighter convergence criterion of 10^{-8}.

4.3.1 Computation

In order to find the optimum level shift value a number of single point calculations on the same CO-heme input geometry used in section 4.2 were set up. The calculations were set to use either B3P86 or B3LYP with the 3-21G* basis set, the number of simulated annealing steps was set to 20, the convergence to 10^{-6} and the vshift (level shift) values from 0 to 1000 millihartrees. The calculations were all run on the same system as specified in section 4.2 using a single processor

for each job. The number of SCF cycles required for convergence at each vshift value were recorded upon completion of each job.

4.3.2 Results and Discussion

The results obtained are presented in the following graph, figure 4.7. Numerical versions of the results and raw data can be found on the support CDRom. In all cases the converged SCF energy was found to be the same implying that the application of level shifting merely affects the convergence efficiency and not the result.

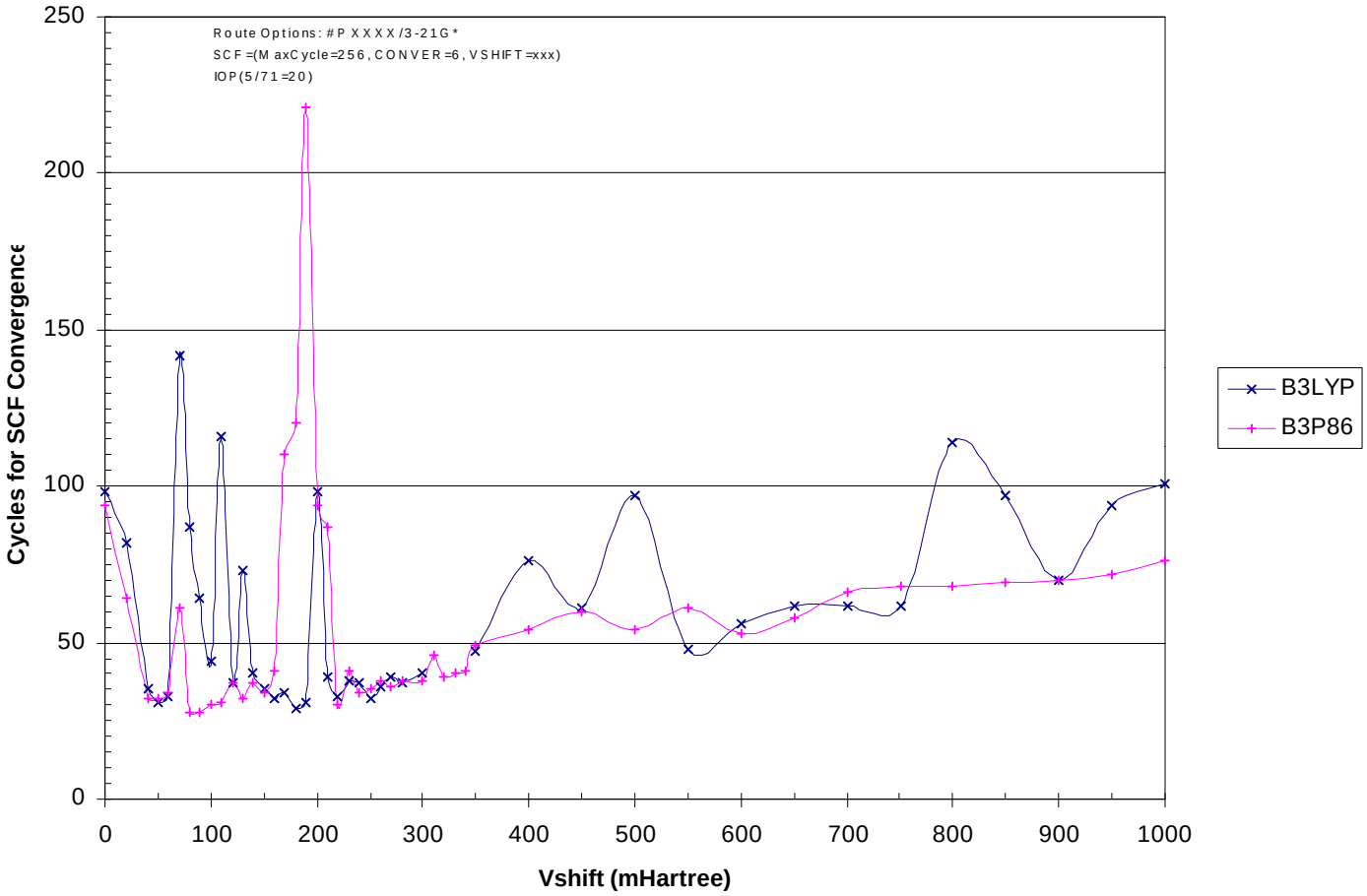


Figure 4.7: Plot showing the number of SCF Cycles required for a single point energy minimisation of CO-heme for B3P86 and B3LYP against the degree of level shifting used.

From the results it can be seen that low level shift values lead to very erratic behaviour. Thus while some of the lowest number of cycles required occur at small vshift values the smallest of changes can dramatically worsen the convergence. Since the effect each vshift value has on the convergence is dependent on the geometry it was concluded that low vshift values, while offering quick convergence for the starting geometries, may suddenly become very poor during the course of an optimisation.

Similarly large vshift values also pose a problem since as expected the trend is towards slower and slower convergence as the degree of level shifting increases. The results are, however, much less erratic above 220 millihartrees. The optimum vshift values show that contrary to the results reported in section 4.2 which were only for a vshift of 500, B3LYP can perform as efficiently as B3P86 if the correct shift is chosen. B3P86, however, appears to be much more tolerant of the vshift value with far less oscillation past 220 millihartrees. Below this value B3P86 is as erratic as B3LYP.

In conclusion therefore a value of 280 millihartrees was chosen to be the optimum value to use since it lies in the centre of the large plateau that occurs between 220 and 350 millihartrees. This was thought to be best as it offers the largest degree of tolerance for different structures and so allows for good convergence throughout the entire optimisation process. Further calculations, not reported here but available as raw data on the CDRom, later showed this to be the case for carbon monoxy and oxyheme systems but that a larger value of 500 millihartrees was required to achieve convergence for deoxyheme systems.

4.4 Optimisation of Annealing Steps

The use of simulated annealing in calculations is often used to improve SCF convergence by allowing more flexibility in the initial wavefunction[86]. The initial temperature for each SCF run is set to be large, (2000 K ~ 3000 K). This adds a larger degree of flexibility to the wavefunction allowing the system to move between various minima. The temperature is then slowly decreased trapping the system in a minima and ultimately leading to convergence. If the cooling process is conducted infinitely slowly, which implies an infinite run time, the trapped minima will be the global minima. In practice, however, infinite runtimes are not viable and the number of steps over which the temperature is reduced has to be specified. In *Gaussian 98* the default is 10 SCF cycles.

To test whether the number of simulated annealing steps would have an effect on the SCF convergence a study similar to those discussed in the above sections was conducted.

4.4.1 Computation

A number of calculations were set up using the same CO-heme geometry as specified in section 4.2. Single point energy minimisations using B3P86/3-21G* were run with the convergence criteria set to 10^{-6} and the vshift, based on the conclusions of section 4.3, set to 280 millihartrees. A total of 32 calculations were run in single processor mode on a Silicon Graphics ONYX 2. The number of simulated annealing steps (IOP 5/71 = XX) was changed for each calculation and

ranged from 0 to 50. The number of SCF cycles required for convergence was recorded at the completion of each job.

4.4.2 Results and Discussion

The results obtained are presented in the following graph, figure 4.8. Numerical versions of the results and raw data are available on the support CDRom. In all cases the converged SCF energy was found to be the same indicating that the number of simulated annealing steps does not affect the results obtained.

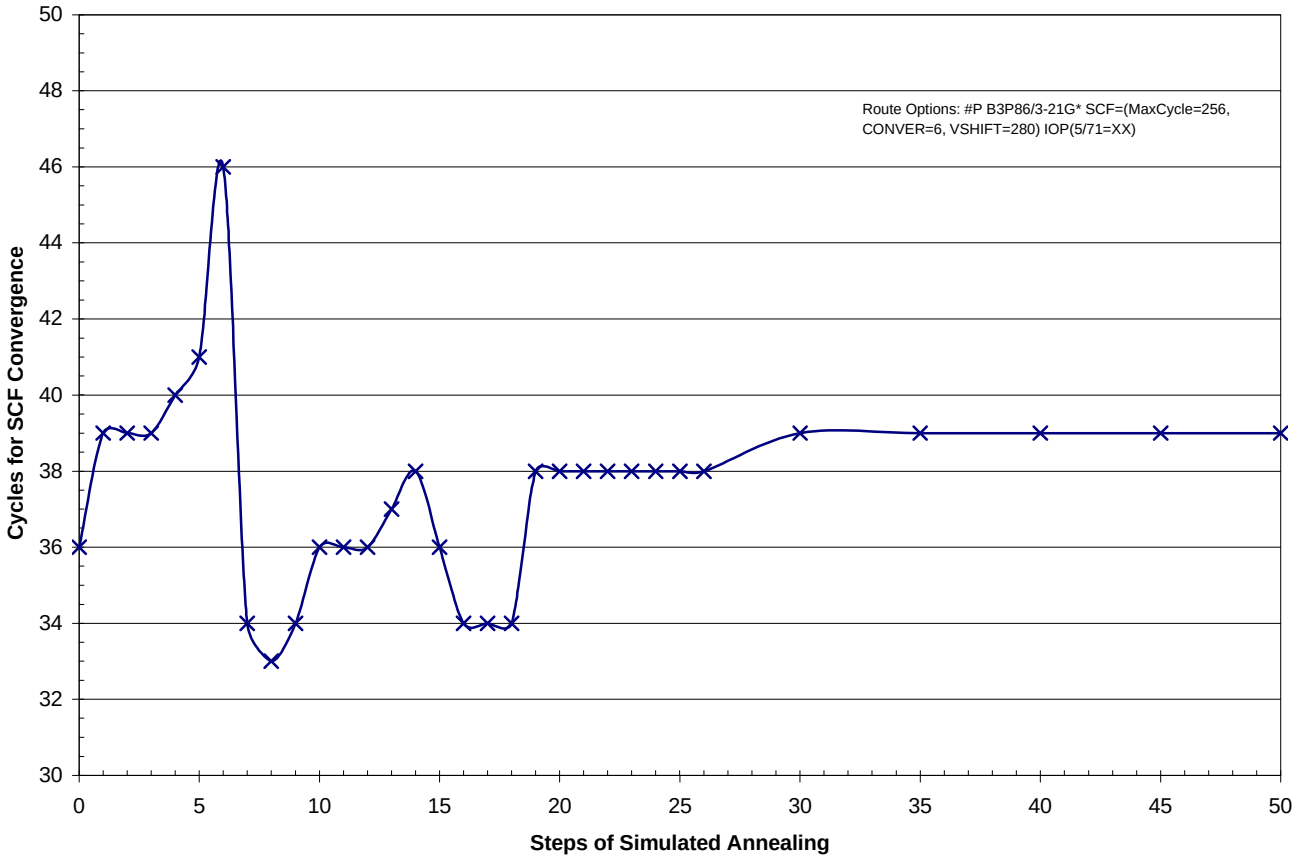


Figure 4.8: Plot showing the number of SCF Cycles required for a single point energy minimisation of CO-heme for B3P86 against the number of simulated annealing steps used.

From the graph it can be seen that the number of simulated annealing steps does indeed have an effect on the rate at which convergence occurs. However, the overall effect is less than that of level shifting making the choice of the number of annealing steps less critical. Low values seem to have the most pronounced effect, oscillating between high and low numbers of cycles. As the number of annealing steps increases above 20 the effect appears to stabilise at an intermediate value. This result could, however, be artificial as above approximately 38 cycles the SCF procedure converges before the annealing process has completed.

It was concluded from the above results that a value of 17 steps would offer the most efficient convergence as this lays at the bottom of a fairly wide trough allowing for a small degree of variation each side as the geometry changes.

4.5 Frequency of Incremental Fock Builds

During each SCF routine the *Gaussian 98* optimiser rebuilds the Fock matrix at regular intervals. The default frequency is to build the Fock matrix incrementally for 20 cycles. During the course of this research, however, it was found that this is not necessarily the optimum setting for calculations where severe convergence problems exist. It was therefore decided to experiment with the incremental fock setting in *Gaussian 98* [IOP(5/37=XX)] to see how this would affect convergence.

4.5.1 Computation

In order to investigate the influence that the frequency of the incremental fock builds would have on the speed of convergence single point energy calculations were set up using the B3P86 functional with the 3-21G* basis set. In order to reduce the time required to obtain results the heme unit used in the previous calculations was reduced to the simplified iron-protoporhyrin IX (*figure 4.9*).

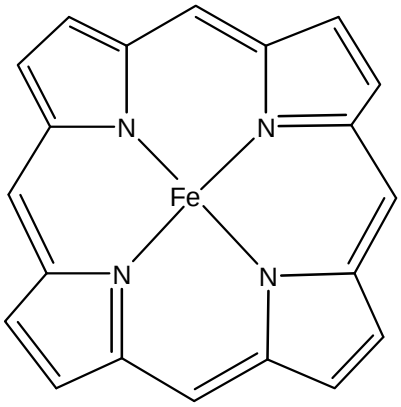


Figure 4.9: 2D schematic showing the structure used as an approximation to heme for the incremental fock investigation.

In all calculations the number of simulated annealing steps was set to 17, the convergence to 10^{-6} and the vshift to 280 millihartrees. The incremental fock build frequency was changed for each calculation with values of 20, 50, 100 cycles and no incremental fock (i.e. ∞ cycles) builds. In all cases the charge was set to zero and the spin state to a singlet. The calculations were run in parallel using 4 processors of the College Computing Services SGI ONYX 2. The electronic energy reported for each SCF cycle was recorded along with the total number of SCF cycles required for convergence with each incremental fock setting.

4.5.2 Results and Discussion

The results obtained are presented in the following graph, figure 4.10 which illustrates the reported electronic energy against SCF cycle. Numerical versions of the results and raw data can be found on the support CDRom.

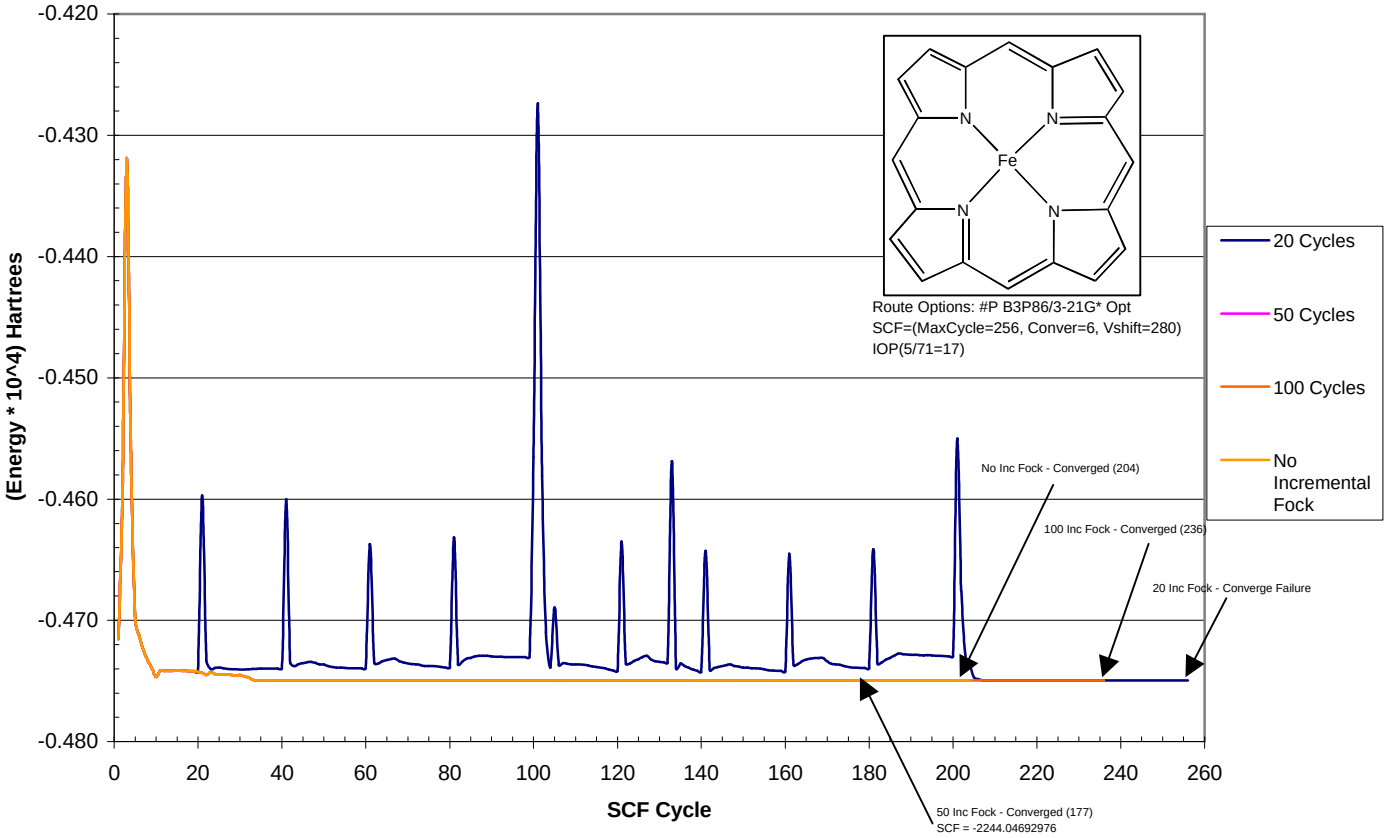


Figure 4.10: Plot showing the electronic energy as a function of SCF Cycle for iron protoporphyrin IX for a number of different incremental fock settings.

It is immediately obvious from the above plot that the default setting of 20 for the frequency of incremental fock builds is actually very poor for this type of iron containing system. The reported energy spikes more or less every 20 cycles as the fock matrix is rebuilt so leading to the energy remaining too high for convergence. The energy eventually reaches approximately the same level as the higher settings but due to the fact that it took so long to get there the calculation fails to converge within the cut off of 256 SCF Cycles. Increasing the number of cycles between fock builds dramatically improves the situation with the energy spikes being eliminated completely. A setting of 50 cycles proved the best, converging after 177 SCF cycles, shortly followed by the calculation with the incremental fock builder switched off (204 SCF cycles).

Thus it was concluded that the best setting to use would be a frequency of 50 cycles. This gives a long enough delay for the energy to initially decrease sufficiently to avoid the spikes but is sufficiently short to avoid lengthening the convergence time by slowing the convergence rate.

4.6 Selection of Spin State

In all calculations to this point the systems studied have always been specified as singlets. This decision was taken since it was the efficiency of convergence rather than the actual converged energy that was of interest and thus the use of restricted formalism in the calculations led to reasonably fast computation. However, when looking at the binding energies it is the converged energy that is of interest hence it was essential that the correct spin state be selected for each system. The presence of iron in the system makes this decision a difficult one since iron can exist in a number of different spin states that are all of similar energy.

Experimentally deoxyheme and deoxymyoglobin are known to be quintets[87,88] unfortunately evidence for the spin state of oxyheme and carbon monoxyheme could not be found. Thus it was decided to try single point calculations on each of the heme starting structures to see which spin state yielded the lowest energy in each case.

4.6.1 Computation

A number of single point calculations were set up for deoxyheme, carbon monoxyheme and oxyheme. The B3P86 functional was used with the 3-21G* basis set. Based on the convergence optimisation calculations reported above the vshift was set to 280 millihartrees, the number of annealing steps to 17 and the incremental fock build frequency to 50 cycles. In all cases the convergence limit was set to 10^{-6}.

The initial structures were taken from the crystallographic data for deoxy horse heart myoglobin provided by the Brookhaven protein databank (1WLA). The desired units were then extracted from the PDB file, protonated manually at standard values for the bond lengths and angles and then the hydrogens optimised using HF/STO-3G while keeping the rest of the structure fixed. The CO and O_2 ligands were then manually built in at the crystallographic distances for sperm whale myoglobin. While not necessarily correct for horse heart myoglobin these structures are sufficient to investigate the spin configuration of each system.

The spin states tested were singlet, triplet, quintet and septet. In all cases the system charge was specified as –2.

4.6.2 Results and Discussion

The results obtained are presented in the following table (*table 4.1*) the *Gaussian* output files are available on the support CDRom.

SCF Energies in hartrees	CO	O_2	Deoxy
Singlet	-3202.78375805	-3233.29233102	-3083.72704993
Triplet	-3202.74751282	-3233.34490619	-3083.79196931
Quintet	-3202.69035664	-3233.31091831	-3083.79948973
Septet	-3202.61962910	-3233.30872941	-3083.69101452

Table 4.1: Results of single point calculations on CO-heme, O_2-heme and deoxyheme for various different spin states.

From the above table it can be seen that the lowest energy configurations are as follows:

CO-Heme *Singlet*
O_2-Heme *Triplet*
Deoxyheme *Quintet*

Based on these results and previous abandoned binding energy calculations, not reported here, it was decided to use these spin states for the binding energy investigations reported in section 5 of this report. The results above are encouraging since the deoxy system concurs with the experimental results. All carbon monoxy systems were thus run as singlets using restricted formalism, while oxy systems were run as triplets and deoxy systems as quintets using unrestricted formalism.

4.7 Basis Set Selection

While conducting the calculations described in the above sections (*4.1-4.6*) and from personal correspondence with various researchers in the field it became apparent that a reasonably large basis set was required to obtain satisfactory results. The use of minimal basis sets such as STO-3G generally led to very poor results or failed convergence. The use of larger basis sets, especially those containing polarisation functions generally converged more efficiently and led to less distorted structures. Thus based on these experiences and personal correspondence[89] it was concluded that a large basis set such as 6-31G* would be required in order to accurately model the heme physics.

The use of such a large basis set, however, had it problems since test calculations showed that the optimisations would likely take a minimum of 6 weeks each with the available resources. A solution was available in the form of hybrid basis sets. Griffiths had reported considerable success

with the use of such basis sets for the study of iron based catalysts[90] using B3P86. From correspondence with Dr. Griffiths and from some initial test calculations it was concluded that the best compromise would be to use 6-31G* for the iron, pyrrole nitrogens and the ligand, 3-21G* for everything else except the hydrogens and STO-3G for the hydrogens. In this way it was ensured that at the point of interest, the iron active site, a large flexible basis set would be used. The hydrogens around the edges of the heme unit were expected to have little influence on the overall heme physics hence the use of a minimal basis set here offered a speed improvement without adversely affecting the results.

This basis set methodology was used for all subsequent calculations.

4.8 Summary

Based on the conclusions made in sections 4.1 to 4.7 above a methodology was arrived at that would allow the binding of oxygen and carbon monoxide to the myoglobin active site to be investigated in a thorough way. The initial problems encountered with convergence were overcome by carefully tuning each of the available parameters discussed above. Therefore using the following methodology (*table 4.2*) it was possible to carry out optimisations at the much tighter convergence limit of 10^{-8} with a timescale that was compatible with that of this research project.

Methodology	
DFT Functional	B3P86
Maximum SCF Cycles	512 cycles
Convergence Limit	10^{-8}
Vshift	280 mhartrees
Simulated Annealing Steps	17 steps
Incremental Fock Frequency	50 cycles
Spin State	Carbon monoxy systems = *Singlet* Oxy systems = *Triplet* Deoxy systems = *Quintet*
Basis Set	Fe, pyrrole N, ligand = 6-31G* H = STO-3G Everything else = 3-21G*

Table 4.2: Summary of the methodology chosen for the binding energy study.

This methodology was used for all the binding energy calculations except for those involving deoxyheme where it was found convergence problems at 10^{-8} still existed. For these systems, as described in section 5 of this report it was found beneficial to increase the vshift from 280 mhartrees to 500 mhartrees. The calculations were also initiated with a looser convergence of 10^{-6}. This was then increased to 10^{-7} and ultimately 10^{-8} after each set of 10 optimisation cycles had been completed. In this way it was possible to study the deoxyheme systems in the same way as the ligated systems without sacrificing accuracy.

In conclusion therefore a working methodology has been developed that solves the convergence problems and thus allows the accurate study of heme containing systems in an acceptable time frame. Solving the convergence problems was an essential goal of this project that has been fulfilled. Without a working methodology for study an investigation of the reactivity of myoglobin would have proven impossible.

5. Ligand Binding Investigations

Once the convergence problems had been solved and a working methodology for study (*§ 4*) developed it was possible to proceed with the main aim of the research project – the investigation of the reactivity of myoglobin. It was decided that the best way to investigate the reactivity of myoglobin would be to look at the predicted binding energies for carbon monoxide and oxygen to the iron atom in heme.

As discussed in the introduction a number of different theories exist for explaining how myoglobin discriminates between oxygen and carbon monoxide. It was decided that the best way to investigate the legitimacy of the various theories would be to initially set up a control system, iron protoporphyrin IX, and find the predicted binding energies for CO and O_2 to this system. By then introducing the different residues into the calculations the effect of each of these individual residues could be investigated. The initial aim was to look at bare heme, heme+his93, heme+his64 and heme+his93+his64 and then to go on and try ONIOM based hybrid quantum mechanics/molecular mechanics calculations on the entire system. Unfortunately time and resource restraints, coupled with the need to use large basis sets to obtain accurate results, meant that the ONIOM calculations had to be abandoned completely and not all of the pure quantum mechanical calculations could be completed on time. Reported in this chapter, however, are the results obtained up to the deadline along with summaries of where the incomplete calculations had reached.

All the results discussed along with the complete and incomplete *Gaussian 98* files are available on the support CDRom.

5.1 Control System (Iron Protoporphyrin IX)

The first step in investigating how the histidine residues surrounding the heme unit effect the ligand binding energies was to develop a control which could be used as a reference. It was decided that iron protoporphyrin IX would make a good benchmark. Since an extensive search of the literature failed to turn up any experimental data for CO or O_2 binding energies to either heme or myoglobin any calculated binding energies would need a reference to compare against. By using the binding energies calculated for a reference system using exactly the same methodology accurate comparisons could be made even if the binding energies weren't formally correct. Using this method any systematic errors in the calculated binding energies were prevented from influencing the results.

5.1.1 Computation

In order to determine the binding energies of CO and O_2 to iron protoporphyrin IX a number of calculations were performed. A crystal structure for iron protoporphyrin IX could not be found so the structure was drawn manually using standard values for the bond lengths and angles[91] and constrained to have d_{4h} symmetry (*figure 5.1*).

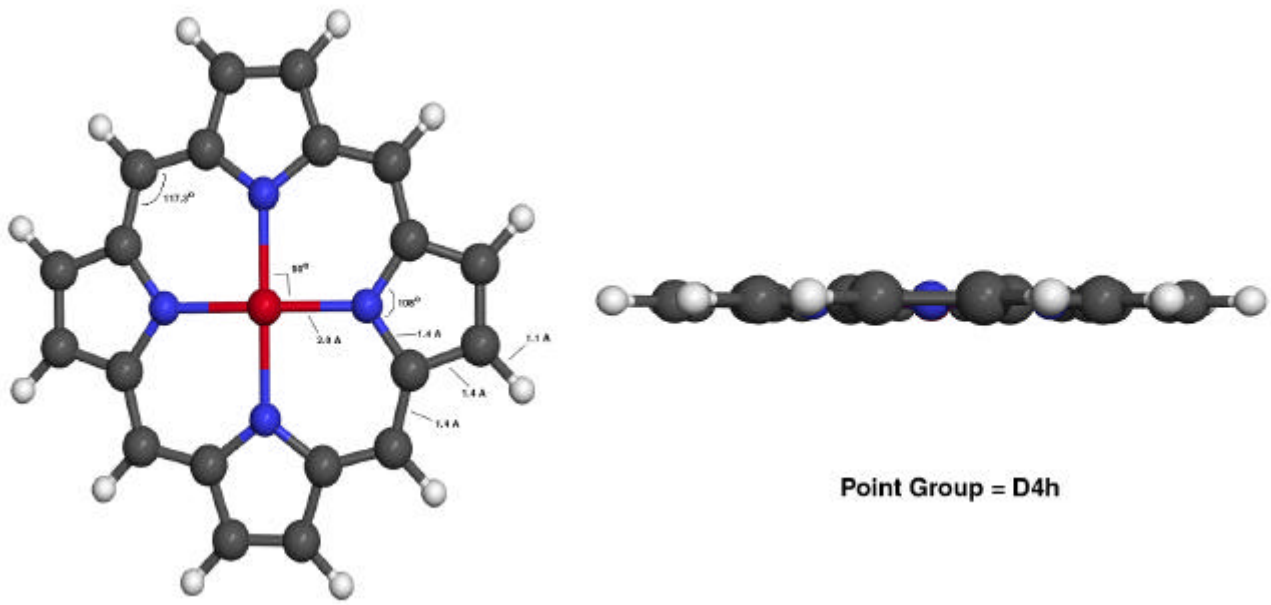


Figure 5.1: Illustration showing bond lengths and angles for the iron protoporphyrin IX input structure.

This porphyrin system is the same as that used by Jewsbury *et al*. in their 1994 paper[92] on CO binding to heme. In this paper they used the porphyrin ring to represent the heme unit and locked this system in d_{4h} symmetry in order to simplify the calculations. It was therefore thought that locking the above system in d_{4h} symmetry would be a reasonable approximation.

Using the deoxy system in figure 5.1 above carbon monoxy and oxy input geometries were constructed by manually adding CO and O_2 at distances typical for this type of system (*figure 5.2*).

Basis Set Superposition Error (BSSE) corrected binding energies were then calculated using the methods highlighted in section 2.7 and the methodology discussed in section 4. In total seven calculations were performed for each system (CO and O_2). The deoxy, carbon monoxy and oxy systems were optimised using B3P86 with 6-31G* on the iron, pyrrole nitrogens and ligands, 3-21G* on everything else except the hydrogens and STO-3G on the hydrogens. Carbon monoxide and oxygen were then optimised separately using B3P86/6-31G*. In all these calculations the convergence criterion was left at the default of 10^{-8}. For the deoxy system the spin state was

specified as a quintet, for the O_2 system a triplet and the CO system a singlet. In all cases the charge was set to zero.

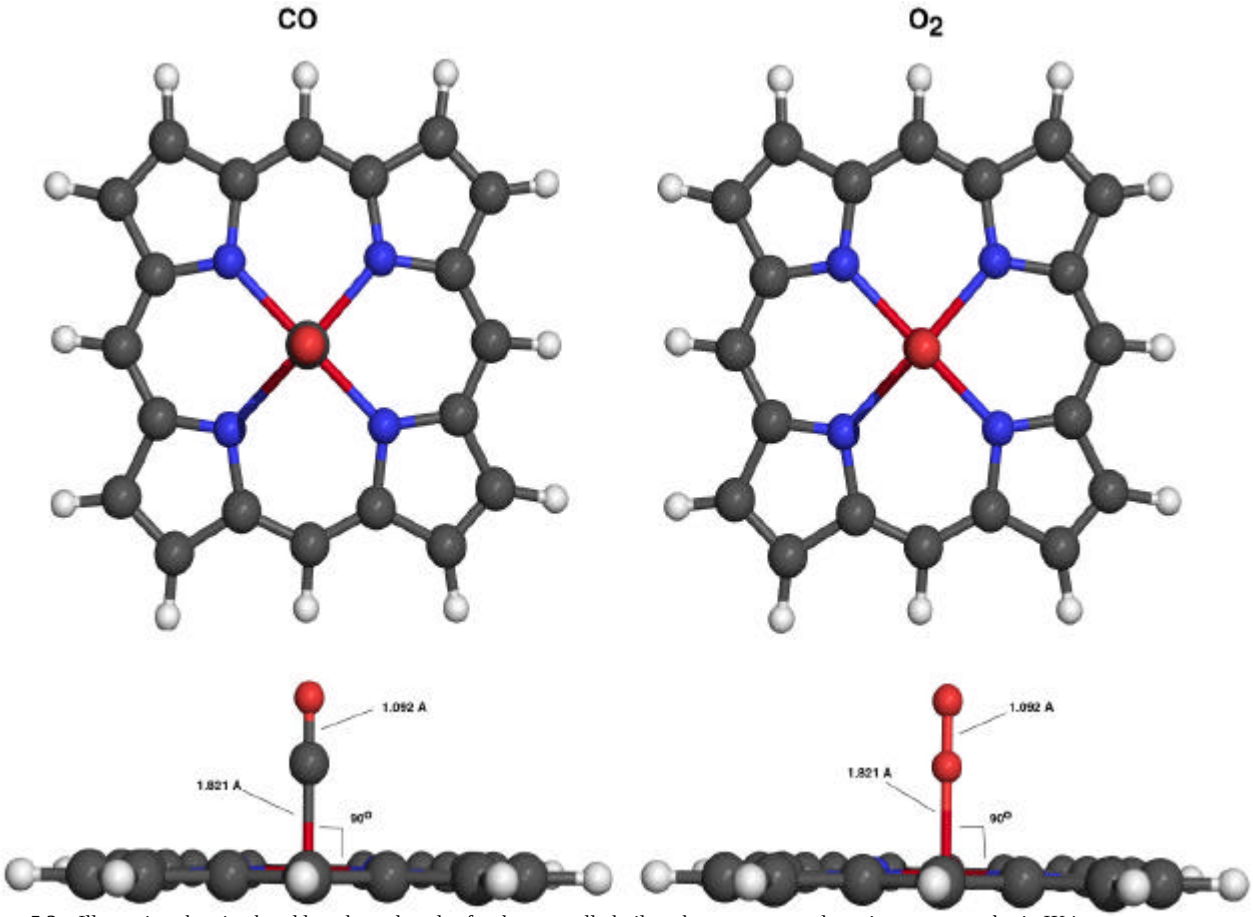


Figure 5.2: Illustration showing bond lengths and angles for the manually built carbon monoxy and oxy iron protoporphyrin IX input structures.

The optimised geometries from the CO-porphyrin and O_2-porphyrin calculations were then extracted from the output files and used for the BSSE corrections. Here single point energy calculations were run using the same methodology as above but with a convergence limit of 10^{-6}. The *Gaussian 98* massage keyword was used to introduce the ghost atoms. The 4 single point energy calculations performed for each system (CO and O_2) were:

1) Ligand in complex geometry (porphyrin removed)
2) Porphyrin in complex geometry (ligand removed)
3) Ligand with porphyrin ghost atoms
4) Porphyrin with ligand ghost atoms

The spin state in each case was specified as either quintet, triplet or singlet depending on the part of the molecule that was left active during the calculation.

From these calculations BSSE corrected binding energies were calculated as described in section 2.7.

All calculations were performed using *Gaussian 98 Revision A.7* on a dual Intel Celeron PC running Redhat Linux 6.0 as described in appendix A. For all calculations except those involving deoxyporphyrin the vshift was set to 280 millihartrees, the number of annealing steps to 17 and the incremental fock frequency to 50 cycles. For the deoxy system it was found that a higher vshift of 500 millihartrees gave more reliable convergence.

5.1.2 Input Structure Test Calculation

In order to test that the input structure would not bias the results a test calculation was set up for carbon monoxyporphyrin using exactly the same parameters as those specified above. However, instead of using linear CO for the input structure the CO was set to be bent with an Fe-C-O angle of approximately 55^{o}. The calculation was then run and found to give exactly the same results as the linear CO job. It was therefore concluded that the initial input geometry for CO and O_2 would not significantly influence the results. Further details are available on the support CDRom.

5.1.3 Results and Discussion

The first important point found in the process of conducting these calculations was that locking the deoxy system in d_{4h} symmetry is not a valid approximation. The system failed to reach an optimised structure when a horizontal mirror plane was present but instead just began to oscillate. By pausing the calculation and then resuming with symmetry turned off (NoSymm) convergence was achieved. The system, however, on first inspection would appear to still possess d_{4h} symmetry. On closer inspection, however, it can be seen that the iron atom has actually been moved out of the plane of the ring by a very small amount (≈ 0.01 Å). While not necessarily significant on its own this small movement would suggest that locking the porphyrin plane in d_{4h} symmetry is not a valid approximation and thus throws doubt on the results reported by Jewsbury *et al*.

Once all calculations had finally been completed the following results, energies and structures, were obtained (*table 5.1*). For full 3 dimensional structures please refer to the structures section of the support CDRom.

5.1.3.1 Deoxy iron protoporhyrin IX

SCF Energy = -2251.20398874 Hartrees

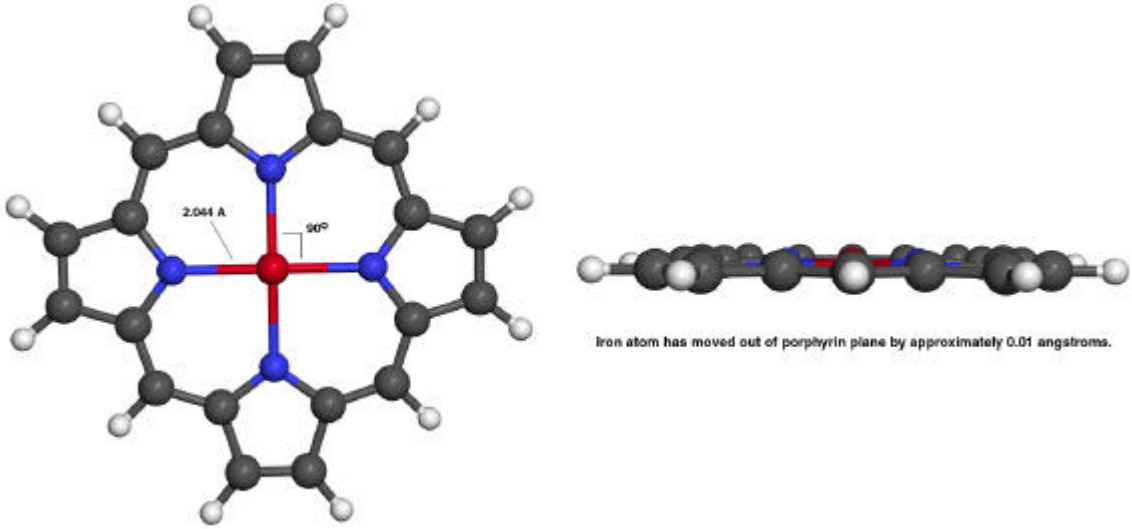


Figure 5.3: Illustration showing bond lengths and angles for the optimised deoxy iron protoporphyrin IX structure.

5.1.3.2 Carbon monoxy iron protoporhyrin IX

SCF Energy = -2364.7901645 Hartrees

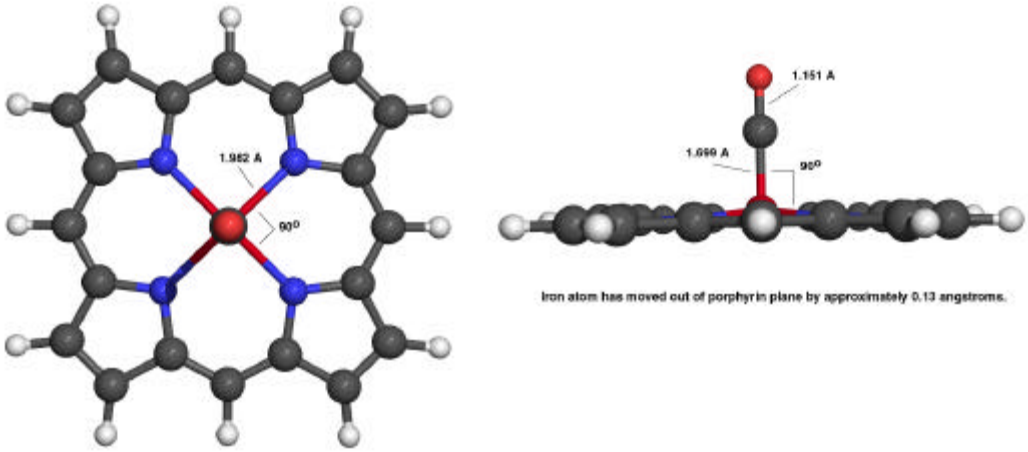


Figure 5.4: Illustration showing bond lengths and angles for the optimised carbon monoxy iron protoporphyrin IX structure.

5.1.3.3 Oxy iron protoporhyrin IX

SCF Energy = -2401.80444808 Hartrees

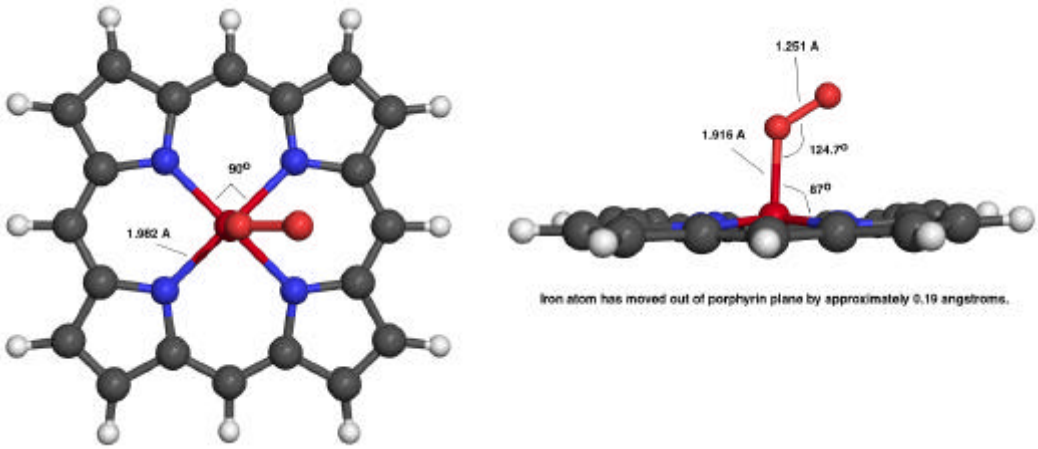


Figure 5.5: Illustration showing bond lengths and angles for the optimised oxy iron protoporphyrin IX structure.

From the images above it can be seen that the porphyrin system itself has changed little in structure during the optimisation. The main changes have occurred to the ligand especially with oxygen. The calculations predict linear CO and bent O_2 which is in line with that suggested by various experimental studies conducted on heme and picket fence porphyrins[93].

The Fe-C bond length in carbon monoxy porphyrin has shrunk from 1.821 Å to 1.698 Å suggesting a relatively strong bond. The C-O bond length has increased from 1.092 Å to 1.152 Å (compared with a C-O length of 1.120 Å for gas phase CO) suggesting that interaction between the iron atom and the carbon atom result in a weakening of the C-O triple bond. From simple consideration of the occupation of molecular orbitals it is apparent that this is indeed what would be expected. The π back bonding from the iron results in electrons being inserted into the π^* orbital of the carbon which while strengthening the Fe-C bond will weaken the C-O bond.

The iron atom has moved out of the plane of the ring by approximately 0.13 Å.

In the oxy porphyrin system the Fe-O bond length has increased from 1.821 Å to 1.916 Å suggesting that the Fe-O bond is very weak. This is conducive with a small binding energy. The O-O bond length also increased from 1.217 Å to 1.251 Å. The iron atom has moved out of the plane of the ring by about 0.19 Å.

In both cases the porphyrin ring remained more or less static with the bond lengths largely unchanged.

5.1.4 Binding Energies and BSSE Correction

The *uncorrected* binding energies can be calculated from the reported SCF energies given above and those calculated for CO and O_2, -113.546644057 hartrees and –150.593310405 hartrees respectively.

This yields the following *uncorrected* binding energies:-

CO binding to Iron protoporphyrin IX	-0.039532 hartrees = -103.79 kJmol^{-1}
O_2 binding to Iron protoporphyrin IX	-0.0071496 hartrees = -18.771 kJmol^{-1}

Correcting these energies for basis set superposition error (*§ 2.7*) yields the following *corrected* binding energies (see support CDRom for tabulated data from each single point calculation).

CO binding to Iron protoporphyrin IX	-0.0367818 hartrees = -96.57 $kJmol^{-1}$
O_2 binding to Iron protoporphyrin IX	+0.0024384 hartrees = +6.40202 $kJmol^{-1}$

What is very interesting about these results is that the corrected binding energy for oxygen is actually positive suggesting that the structure is not bound. Carbon monoxide on the other hand is still predicted to be bound with a binding energy of approximately 100 $kJmol^{-1}$. This is a very encouraging result as it is exactly in line with that expected from experimental observations[94] where unhindered porphyrins are observed to have an affinity of approximately 30,000 times more for CO than for O_2. This would suggest that the methodology used is valid and that binding energy comparisons can be reliably made between this porphyrin system and other more complex heme systems.

It was thus concluded from these results that the procedure used yields valid results. It was therefore decided to extend the investigation and use the developed methodology to run calculations on heme and heme + various residues. The predicted binding energies could then be compared with the results discussed above to determine the influence each different residue has on the CO and O_2 binding.

5.2 Carbon Monoxide Binding to Heme Systems

The first step in looking at the way in which myoglobin discriminates between carbon monoxide and oxygen was to see if the binding energy of CO is reduced by any of the residues close to the heme.

Jewsbury *et al*. reported in 1994[95] that calculations using MP2 showed that the proximal histidine was responsible for the reduced CO affinity by reducing the CO binding energy and forcing the CO ligand to bind in an unfavourable bent configuration. Their work, however, involved the use of a large number of simplifications to make the calculations feasible using computers available in 1994. The approximations involved using iron protoporphyrin IX locked in d_{4h} symmetry to represent the heme unit. This has already been shown to be a poor approximation by the results reported in section 5.1 above. The second approximation was to rotate both the distal and proximal

histidines to form an artificial mirror plane. The histidine units were also truncated to just their imidazole components. Finally a Hay and Wadt 16-electron pseudo potential[96] was used for the iron. These approximations throw doubt on the validity of the results obtained and so it was decided to investigate the effect of each histidine unit using the more thorough methodology developed in section 4.

5.2.1 Computation

Calculations were set up using the methodology summarised in section 4.8. Input structures for bare heme, heme+his93, heme+his64 and heme+his93+his64 were extracted from the crystallographic data for deoxy horse heart myoglobin provided by the Brookhaven protein databank (1WLA). To simplify the calculations the histidine units were truncated at the β-carbon. The desired units were then manually protonated at standard values for the bond lengths and angles and then the hydrogens optimised using HF/STO-3G while keeping the rest of the structure fixed. These structures (*figure 5*.6) were then used for the deoxy calculations with the charge set to –2 and the spin state, based on the results reported in section 4.6, to a quintet.

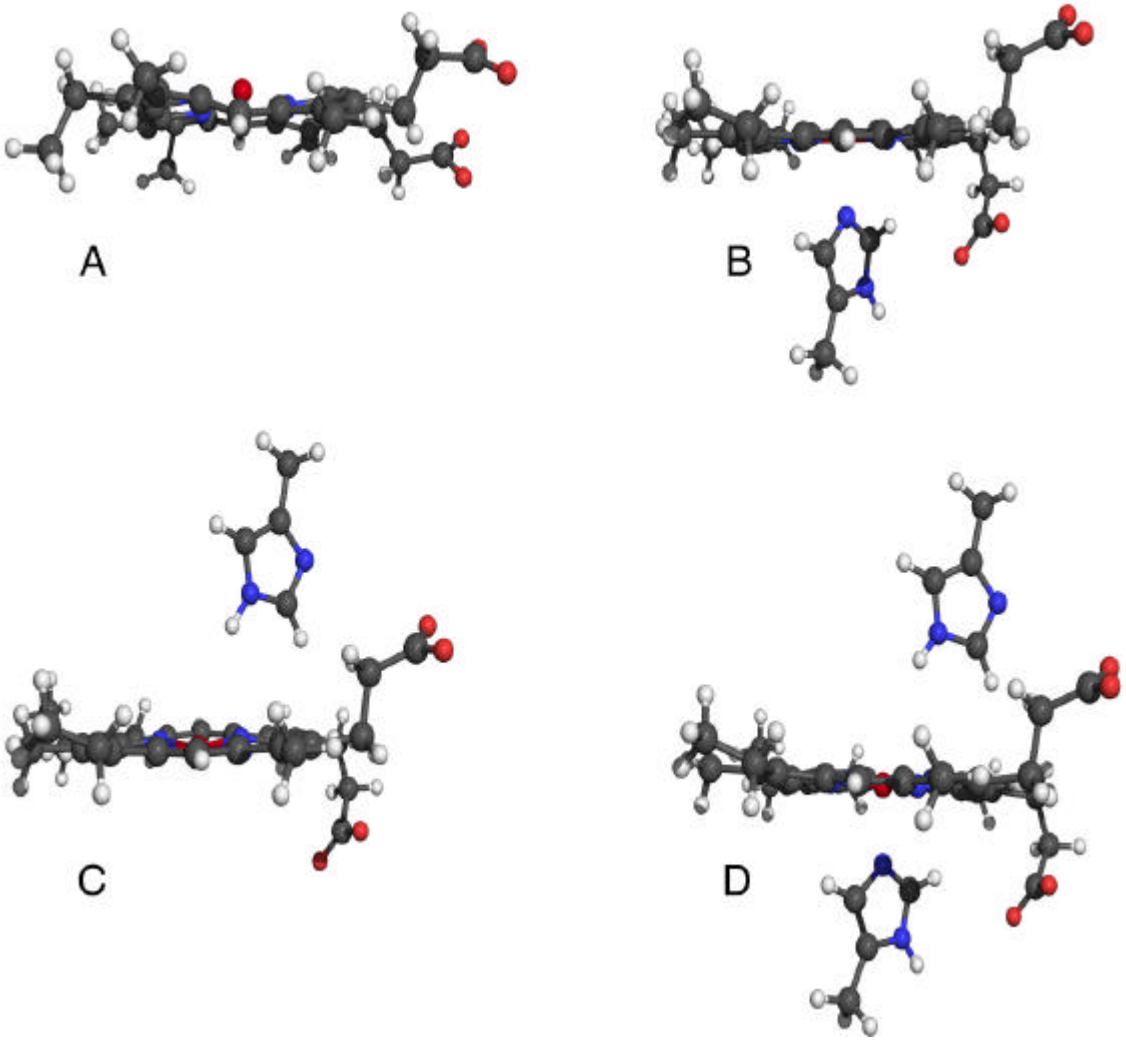


Figure 5.6: Illustration showing the deoxy input structures used. A – Deoxyheme, B – Deoxyheme+his93, C – Deoxyheme+his64, D – Deoxyheme+his93+his64.

For the carbon monoxy structures the carbon monoxide ligand was built in manually with a linear geometry, perpendicular to the heme plane. The Fe-C bond length was set to 1.821 Å and the C-O bond length to 1.152 Å. The charge was set to –2 and the spin state to a singlet. The input files used are available on the support CDRom.

These structures were then optimised using B3P86 with 6-31G* on the iron, pyrrole nitrogen and CO ligand, 3-21G* on everything else except the hydrogens and STO-3G on the hydrogens. The number of annealing steps was set to 17 and the incremental fock frequency to 50 cycles in all cases. For the CO structures the vshift was set to 280 mhartrees and the convergence to 10^{-8}. For the deoxy structures a larger vshift of 500 mhartrees was required to obtain convergence. It was also necessary to start the deoxy calculations off at a convergence of 10^{-6} this was then increased to 10^{-7} after approximately 10 optimisation steps and then finally to 10^{-8} after a further 10 steps.

The optimised structures and predicted SCF energies were recorded at the end of each optimisation. The optimised CO-heme system geometry reported was then extracted and used to calculate BSSE corrected binding energies as outlined in section 2.7.

5.2.2 CO Binding to Bare Heme – Results and Discussion

The carbon monoxyheme calculations finished relatively quickly yielding the following optimised structure (*figure 5.7*).

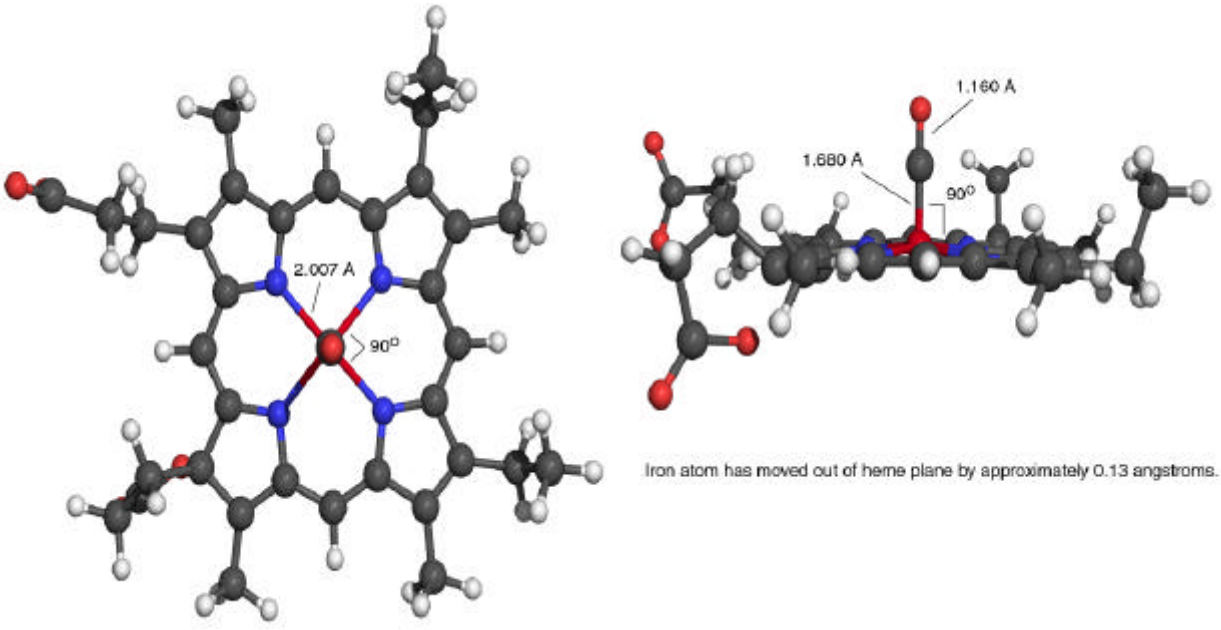


Figure 5.7: Illustration showing the B3P86 optimised carbon monoxyheme structure.

The structure obtained is similar to that found for the control system (iron protoporphyrin IX) the CO ligand has remained linear and perpendicular to the heme plane as expected with the Fe-C bond length shrinking from 1.821 Å to 1.680 Å implying a strong bond. At the same time the C-O bond length has lengthened, due to π back bonding, to 1.160 Å. The iron atom has also moved out of the plane of the ring by approximately 0.13 Å in exactly the same fashion as that observed for the control system.

The deoxyheme calculations took considerably longer to finish due to poor convergence. Using a stepped procedure for the convergence criteria solved this problem, however, leading to the optimised structure shown below (*figure 5.8*).

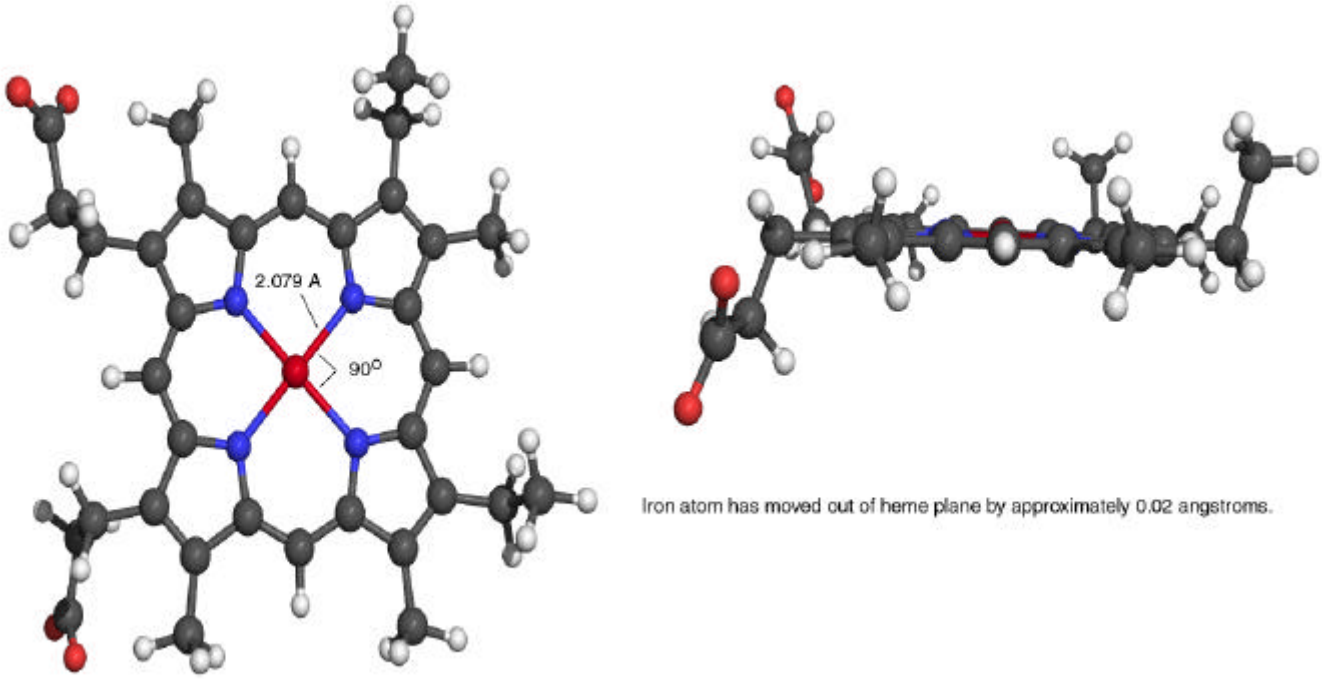


Figure 5.8: Illustration showing the B3P86 optimised deoxyheme structure.

For the deoxy system the ring has moved to a flat configuration but with the iron atom still out of the heme plane by a small amount (≈ 0.02 Å). The iron nitrogen distances have remained approximately the same at 2.051 Å and stayed symmetrical. Most of the structural changes have occurred with the carboxylate side chains that have swivelled away from each other. An animation showing exactly how the structure changed during optimisation is available on the support CDRom.

5.2.2.1 Binding Energies and BSSE Correction

The *uncorrected* binding energy calculated from the SCF energies reported for CO-heme (-3210.27403972 hartrees), deoxyheme (-3096.68842427 hartrees) and CO (-113.546644057 hartrees) is:

Uncorrected CO binding to bare heme = -0.038971 hartrees = -102.32 $kJmol^{-1}$

Correcting this for basis set superposition error (*section 2.7*) yields a *corrected* binding energy of:

Corrected CO binding to bare heme = -0.0393878 hartrees = -103.41 $kJmol^{-1}$

This is very similar to the binding energy calculated for iron protoporphyrin IX (-96.57 $kJmol^{-1}$) which, as expected, implies that the side chains on the heme unit (4 methyl, 2 ethyl and 2 propanoate side chains) have more or less no effect on the binding of carbon monoxide to the heme iron atom. They may, however, especially the propanoate side chains, play an important role in anchoring the heme unit within the protein matrix. In conclusion therefore based on these results it would appear that approximating the heme unit to just the porphyrin ring used in section 5.1 is valid. However, as illustrated in the results for heme+his93 discussed in section 5.2.3 it can be seen that ignoring the presence of the side chains may be over simplifying the situation.

5.2.3 CO Binding to Heme+his93 – Results and Discussion

As with CO-heme the carbon monoxy calculations finished relatively quickly with few convergence problems yielding the following optimised structure (*figure 5.9*).

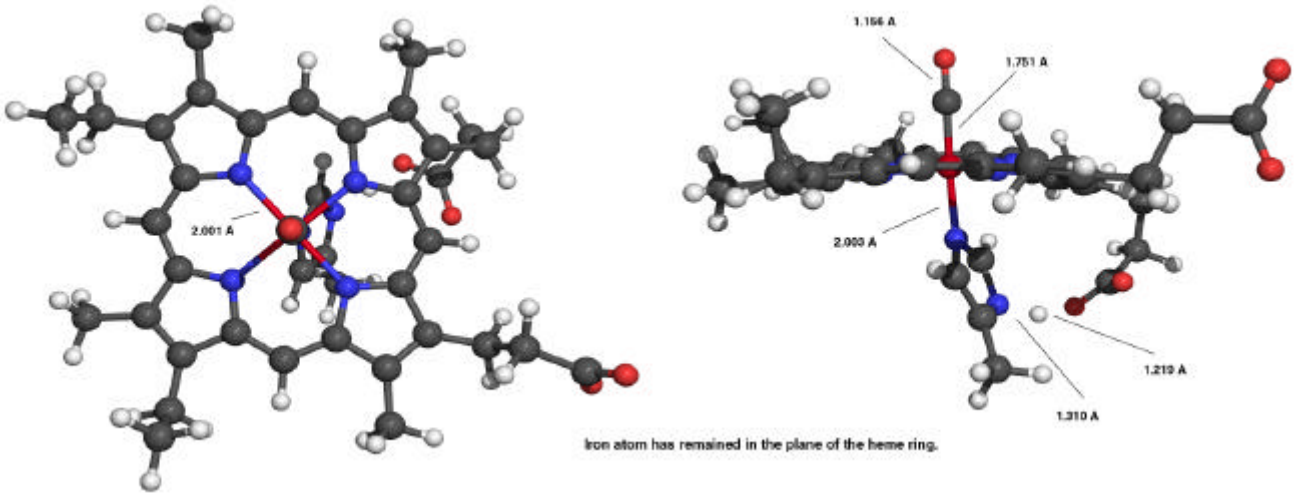


Figure 5.9: Illustration showing the B3P86 optimised CO-heme+his93 structure. The formation of a hydrogen bond is clearly visible on the right hand side image.

There are several interesting points to note about these results. The first is that contrary to what would be expected from the Jewsbury paper the CO ligand is still perfectly straight and perpendicular to the heme plane. The Fe-C bond length is slightly longer for this system than the bare heme system with a bond length of 1.751 Å compared with the 1.680 Å predicted for bare heme. This suggests a slightly weaker bond. The C-O bond length is more or less the same as that for the bare heme system at 1.156 Å. This suggests that there is still the same degree of π back bonding between the iron and the carbon and that it is therefore the σ framework which has been reduced in strength.

What is also interesting is that there is a predicted hydrogen bond between the proximal histidine and one of the carboxylate side chains. This has the effect of holding the proximal histidine slightly off centre forcing the iron atom to remain in the plane of heme ring. This is obviously unfavourable for binding CO when compared with the fact that the iron atom when not hindered by the histidine would prefer to be out of the plane of the ring by approximately 0.13 Å.

The deoxy system, as before, took longer to converge than the carbon monoxy system but using the stepped convergence criteria methodology suggested for deoxy systems yielded the following optimised structure (*figure 5.10*).

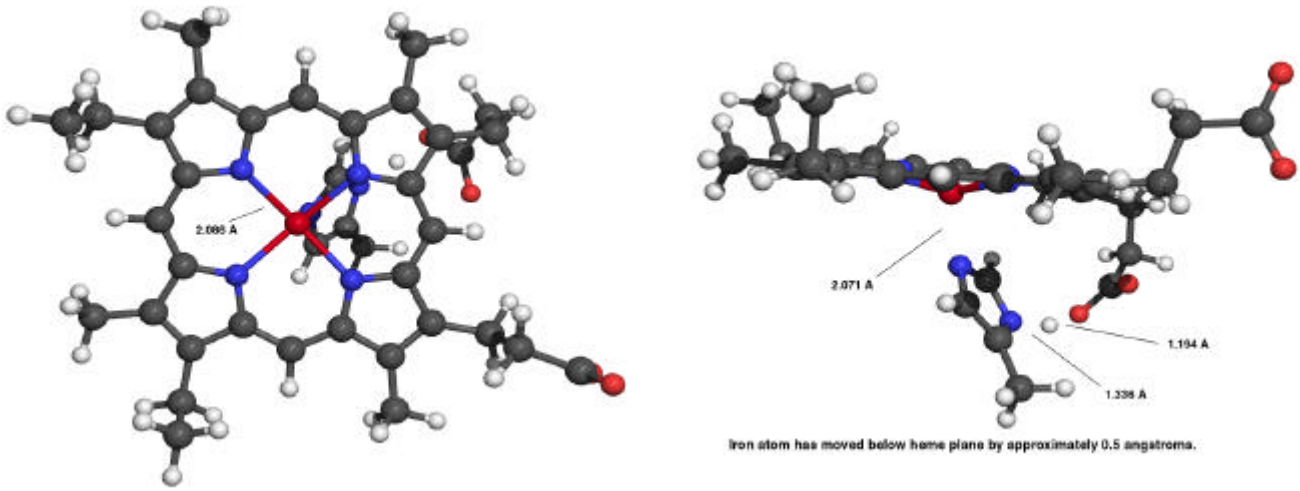


Figure 5.10: Illustration showing the B3P86 optimised Deoxyheme+his93 structure. The formation of a hydrogen bond is clearly visible on the right hand side image.

Again we can see the formation of a hydrogen bond between the proximal histidine and one of the carboxylate side chains. This holds the iron atom below the plane of the ring in this case by approximately 0.5 Å.

What can be concluded from the presence of these hydrogen bonds in both the deoxy and carbon monoxy calculations is that although they may not strictly be present in the protein the possibility of their formation does exist and thus it must be concluded that the heme side chains do play an important role in controlling the position and orientation of the proximal histidine. This in turn controls the position of the iron atom within the heme plane. Thus approximating the heme unit to a simple porphyrin ring, as was made by Jewsbury *et al*. is not a valid assumption.

5.2.3.1 Binding Energies and BSSE Correction

The *uncorrected* binding energy calculated from the SCF energies reported for CO-heme+his93 (-3475.19252668 hartrees), deoxyheme+his93 (-3361.60192298 hartrees) and CO (-113.546644057 hartrees) is:

Uncorrected CO binding to heme+his93 = -0.04396 hartrees = -115.417 $kJmol^{-1}$

Correcting this for basis set superposition error (*section 2.7*) yields a *corrected* binding energy of:

Corrected CO binding to heme+his93 = -0.0332201 hartrees = -87.219 $kJmol^{-1}$

In this case the proximal histidine has resulted in the CO binding energy being reduced by approximately 15 % from that predicted for CO binding to bare heme. The proximal histidine does indeed therefore play a role in reducing the CO binding affinity. This is probably due to competition for the d orbitals of the iron. One possibility is that the iron bound nitrogen of the proximal histidine feeds electrons into the σ^* of the iron and in turn reduces the σ bonding strength between the iron and the CO ligand. This is, however, just speculation and could not be verified based on the results obtained.

What can be concluded from this investigation, however, is that while the proximal histidine reduces the CO binding affinity its effect is really only minimal and cannot solely account for the huge change in relative affinity for CO and O_2 observed experimentally for bare heme and myoglobin. This goes against the results suggested by Jewsbury *et al*. where it was suggested that the change in binding affinity is due almost entirely to the presence of the proximal histidine. The results obtained also imply that the proximal histidine is not responsible for the supposedly bent CO geometry in myoglobin. This investigation has been carried out with far less approximation

than that made by Jewsbury *et al.* and therefore one is forced to concluded that the results reported in 1994 are most likely incorrect.

5.3 Abandoned Calculations

5.3.1 CO Binding to Heme+his64 and CO Binding to Heme+his93+his64

From the results obtained for CO-heme and CO-heme+his93 the next step was to look at CO-heme+his64 to see what effect the distal histidine has on the orientation and strength of CO binding. Following this the intention was to look at CO binding to heme with both the proximal and distal histidines present. In this way it would be possible to quantify to what extent each histidine unit affects the binding energy and also whether there are any combined effects occurring due to a symbiotic relationship between the two histidine units. Finally from this it would be possible to determine whether the difference in relative binding energies is due to just these two histidine units or to the whole protein.

Unfortunately lack of time, resources and problems getting the calculations involving the distal histidine to converge meant that these calculations had to be abandoned. A major problem observed was that without the protein matrix present in the calculation the distal histidine does not remain in position but either flies off away from the heme system or just oscillates around the heme ring as can be seen by the following image showing the starting structure for CO-heme+his64 and the structure after 108 optimisation steps (*figure 5.11*). Copies of the *Gaussian* input and output files up to the point where the calculations were abandoned are available on the support CDRom.

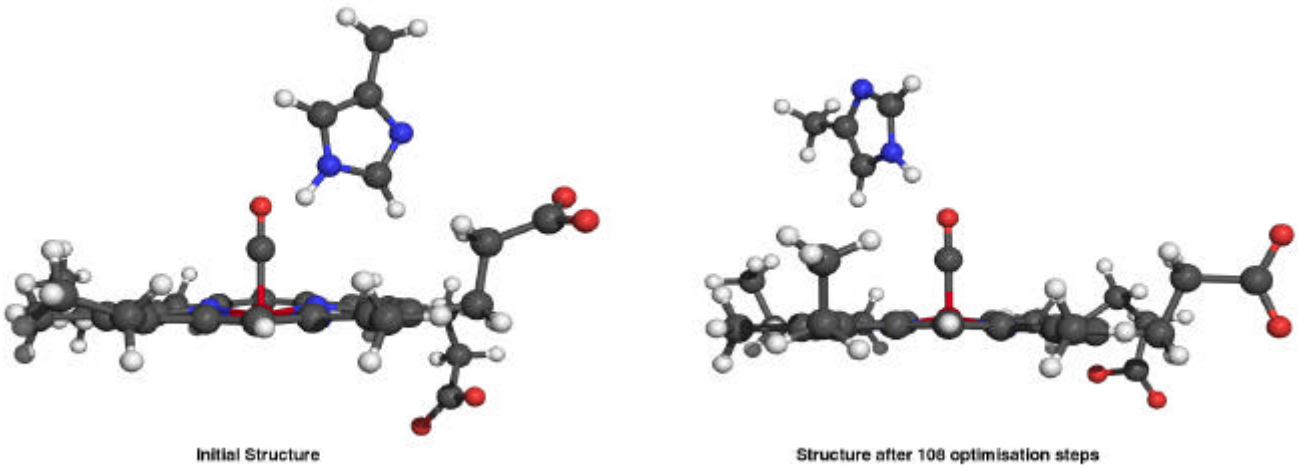


Figure 5.11:- Illustration showing CO-heme+his64 input structure and the predicted structure after 108 optimisation cycles.

It can be seen that the distal histidine has effectively swapped sides of the heme ring. This oscillation meant that convergence of the *ab initio* calculations could not be achieved. This oscillation can be seen very clearly in the animation available on the support CDRom showing how the structure of this molecule changed as the optimisation progressed. Unfortunately the time spent developing a working methodology and overcoming the initial SCF convergence problems meant that time was not available to look at ways for solving this oscillation problem. Therefore the role of the distal histidine (his 64) and the combined role of the proximal (his 93) and distal histidines in influencing the binding energy of CO to the iron atom in myoglobin cannot be determined from the data available. Future study should therefore concentrate on solving this oscillation problem and also on QMMM calculations where the entire protein matrix can be included.

5.3.2 O_2 Binding Energy Calculations

Having first looked at the role of the proximal and distal histidine units in affecting the CO binding affinity the second stage of the investigation was to look at how the oxygen binding affinity is affected by the two histidine units. Initially calculations were set up using exactly the same methodology as that used for the CO binding energy calculations (*§ 5.2.1*). The O_2 ligand was added manually to each of the deoxy systems at an Fe-O distance of 1.821 Å perpendicular to the heme plane, an O-O distance of 1.070 Å and a Fe-O-O bond angle of 110.9^{o}.

The intention was to follow the same procedure as that used for the CO binding investigations. However, the deadline for this project and the available resources meant that none of the calculations could be completed on time. The O_2-heme calculation was abandoned after 62 steps while the O_2 heme+his93 calculation had to be abandoned after 49 optimisation steps. Unfortunately the predicted SCF energy in both cases was still fluctuating in the 2nd decimal place making an accurate prediction of the final optimised energy almost impossible. The last predicted structure for each calculation is shown in figure 5.12 on the following page.

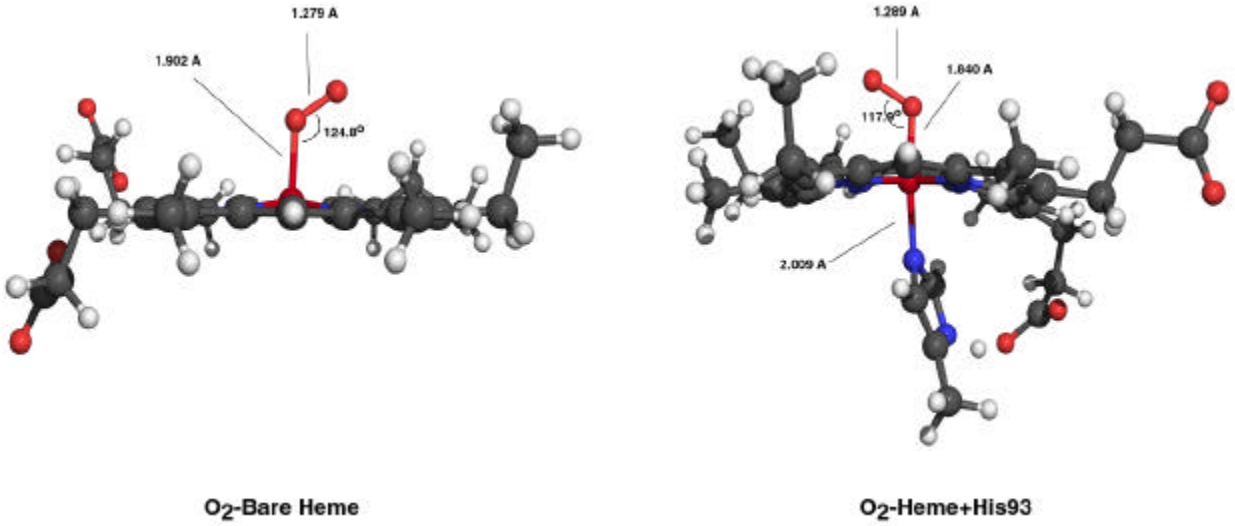


Figure 5.12:- Illustration showing O_2-heme predicted structure after 62 steps and O_2-heme+his93 after 49 steps.

What can be seen from these structures is that for the bare heme we have a very long Fe-O bond of 1.902 Å, very similar to the porphyrin ring system, suggesting a weak bond as expected. The iron atom is also predicted to be raised out of the heme plane as was observed for carbon monoxyheme. We also have a bent O_2 structure as expected.

For O_2-heme+his93 we again have a bent O_2 structure and the prediction of a hydrogen bond as observed for the CO structure. What is interesting here though is that the Fe-O distance is considerably shorter than for the bare heme system at 1.840 Å which suggests the bond between the iron and oxygen ligand may be stronger than the bare heme system for O_2. It may be possible that the O_2 ligand is more π bonded to the iron atom than σ bonded and so is not as effected by σ^* donation from the proximal histidine. However, it must be remembered that these are not optimised structures so conclusions cannot really be made until full optimisations have been run. Due to time restraints this will have to be left to future work.

6. Conclusions and Future Work

Unfortunately the time restraints imposed on this research project and the number of unforeseen difficulties that had to be overcome meant that all of the intended aims could not be completed on time. However, a number of conclusions can be made based on the work that was accomplished.

The first conclusion (*§ 3.1*) is that contrary to what is implied in the *Gaussian 98* manual the FMM routines in *Gaussian 98 Revision A.7* do not offer linear scaling with system size. The scaling is better than density functional methods but is still far from linear. The implementation also does not scale well to multiprocessing systems meaning that large scale *ab initio* calculations on proteins the size of myoglobin are outside of the reach of current technology. Thus future work in this field should be directed towards creating algorithms that offer true linear scaling with acceptably small pre factors. A lot of work also needs to be directed towards making the algorithms scale well to multiprocessor systems since the future of high end computing would seem to be towards increasingly parallel systems.

When starting the initial binding energy calculations a major problem with *ab initio* theory, and density functional in particular[97], when treating transition metal elements was discovered. The large number of available spin states of similar energy for iron meant that getting the calculations to converge was very difficult. Thus in section 4 the development of a methodology for looking at such troublesome compounds was outlined in detail. The most important conclusion of this project is that a methodology has now been developed that allows accurate study of heme systems using density functional theory that does not suffer from the convergence problems initially experienced. There is no reason why this methodology could not be adapted for any other system which poses convergence problems. This was a major objective of this project and one that has been fulfilled successfully.

In section 5 of this report was discussed the binding energy calculations. Here the conclusions drawn are that carbon monoxide does indeed appear to bind to porphyrin systems much more strongly than oxygen. It was concluded that the methodology developed in section 4 did indeed appear to work reliably and application of this to CO-heme and CO-heme+his93 would suggest that the proximal histidine plays a very small part in reducing the CO binding energy (≈15 % reduction) and that linear CO is still the preferred geometry. This is contrary to what was reported by Jewsbury *et al.* and, since this work introduced fewer approximations, throws doubt on their results. Heme+his64 and heme+his93+his64 calculations were abandoned due to a combination of lack of time and problems getting the system to accurately represent that found within the protein matrix.

Thus any future studies in this field should really concentrate on investigating the effects of the distal (his 64) histidine to determine its role in ligand binding. To do this will require an investigation into ways in which the his 64 ligand can be prevented from oscillating around the heme ring. One possibility is to do a partial optimisation with the Fe his 64 distance and dihedral fixed. However, this sort of treatment is likely to lead to convergence problems that will required solving. Calculations should also be carried out to see if a hydrogen bond between the bound O_2 and the distal histidine is indeed present and to predict how this affects the binding affinity for oxygen. This calculation was unfortunately not carried out due to time constraints.

Another field which is worth investigating is hybrid quantum mechanics/molecular mechanics (QMMM) studies on the system to see how the predicted results vary with the ratio of the QM to MM fragment sizes. This was an original aim of this research project and some modification was made to the *ONIOM* implementation in *Gaussian 98* (*appendix c*) to correct certain memory problems and make such studies feasible. These, however, had to be shelved due to resource and time restraints.

In conclusion therefore this project has been a success and shown that these types of *ab initio* calculations can be used to model heme systems. Given more time, a more thorough study would have been possible and future work should offer promising results.

Appendix A - System Benchmark Calculations

During the course of this research it was decided to look at the relative performance of *Gaussian 98* on two different machine architectures. The two machines compared were as follows:

A.1 44 Processor Silicon Graphics ONYX 2

Manufacturer	Silicon Graphics Inc.
Processors	44 x 195 MHz IP27 MIPS R10000 (64 Bit)
Processor Cache	4Mb Level 2
Total Memory	11 Gb (ECC)
Local Shared Memory per Processor	512 Mb (ECC)
Disk Interface	SCSI (Raid)
Operating System	Irix 6.5
Compiler	MIPS Pro F77 v7.2
Approximate Total Cost	£1,000,000.00
Approximate Cost per Processor	£22,800.00

A.2 Dual Intel Celeron

Manufacturer	Intel Processors, ABIT Motherboard, IBM Hard Drives, Generic Memory.
Processors	2 x 466 MHz Intel Celeron Processors (Running at 80 MHz x 7 = 560 MHz)
Processor Cache	128 Kb Level 2 (Full Speed)
Total Memory	256 Mb
Local Shared Memory per Processor	256 Mb
Disk Interface	3 x EIDE UDMA 66 (Software Raid)
Operating System	Redhat Linux 6.0 (Kernel 2.2.14)
Compiler	Portland Group IA-32 F77 v3.1
Approximate Total Cost	£460.00
Approximate Cost per Processor	£230.00

A.3 Benchmark Calculations

A.3.1 Linux Optimisation

The first stage of the comparison was to get *Gaussian 98* running correctly, and in parallel, under Linux as this had not been tried before in the department. The first problem encountered was the very low default maximum shared memory block that is allowed in Linux. This is controlled by the variable SHMMAX in the kernel header file shmparam.h. By default the value of this variable is 0x2000000 which is the hexadecimal representation of 32 Mb. Such a small shared memory size is insufficient for running *Gaussian* jobs which frequently require in excess of 64 Mb. Fortunately this problem was not difficult to fix. Simply editing shmparam.h and changing the value of SHMMAX to 0xFFFFFFF (255 Mb) followed by recompiling the kernel solved the problem. The value of SHMMAX could conceivably have been increased further but as the

machine only has 256 Mb of RAM a maximum shared memory block size of 255 Mb is sufficient.

After the initial problems had been solved it was decided to look at the different compiler options available in order to optimise the performance as much as possible. The level of compiler optimisation was left at the defaults set in the *Gaussian* makefile as modification of these values can lead to errors in the results produced. The two options investigated, therefore, were the compiler cache alignment and the BLAS library used.

A.3.1.1 Compiler Cache Alignment

The Portland group compilers offer the option to align data structures with the cache size of the processors used (-Mvect=cachesize:xxxxxx). Setting this option correctly means that the compiler will try to maximise the reuse of data in the fast processor cache memory rather than the larger, but considerably slower, main memory. The default value is 512 Kb. This is the amount of half speed, off chip, cache on standard Intel Pentium II and Pentium III processors. The Intel Celeron processor, however, is a cheaper version of the Pentium II. Intel saves money on the production of Celeron processors by reducing the size of the level 2 cache to 128 Kb. The cache in a Celeron, however, is based on chip and runs at the same speed as the processors core. Thus the difference in performance between a Celeron and a Pentium II is negligible[A1].

The reduced level 2 cache, however, does mean that the Portland compiler default is no longer correct. Thus it was decided to test the performance of *Gaussian 98* with the data structures aligned at both 512 Kb and 128 Kb.

A.3.1.2 BLAS Libraries

The second optimisation investigated was the BLAS library used in the construction of the *Gaussian 98* utility library. The BLAS library provides all the essential basic linear algebra routines used by *Gaussian 98* and its performance is thus critical to the overall performance of *Gaussian 98*.

Three different BLAS libraries were tried both in single and dual processor mode. The first library tried was the default BLAS routines provided with the *Gaussian 98* distribution. This consists of a generic (processor independent) piece of fortran code which is compiled into the

util library by the compiler. The second library tried was the Pentium II optimised BLAS library provided by the Accelerated Strategic Computing Initiative (ASCI) program at Sandia National Labs[A2]. This library is the same as that used by the current (March 2000) world's fastest supercomputer (Intel ASCI Red) and has been specifically optimised for Pentium II processors. The third library tried was the optimised BLAS library provided with the Portland Group compiler. This library has been optimised specifically for use with the Portland compilers.

A.3.1.3 Procedure

The procedure used was the same as that detailed in section *3.1.1* above. The same linear set of straight chain alkanes was used running from methane to icosane. Geometry optimisations were carried out on the molecular mechanics structures using HF-3-21G with a memory allocation of 6 MW (48 Mb) and the use of symmetry specifically turned off using the NoSymm keyword. The reported CPU time for each job was then extracted from the files, converted to seconds and exported to *Microsoft Excel* as before.

A.3.1.4 Results and Discussion

The first set of results obtained, for the cache alignment value, are given in figure A.1 below:

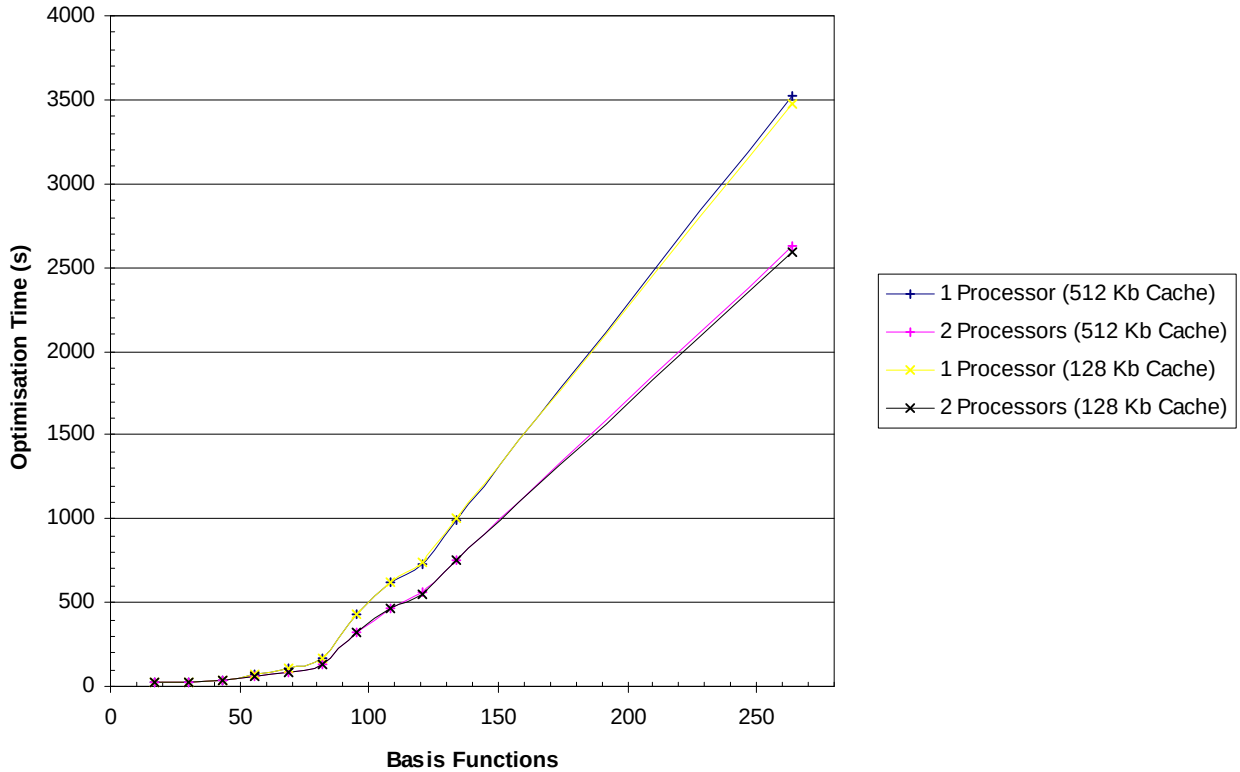


Figure A.1: Plot showing time for geometry optimisation at HF/3-21G as a function of the number of basis functions for a range of straight chain alkanes. The 4 lines correspond to 1 and 2 celeron processors with the data structures aligned at 512 Kb or 128 Kb. The Sandia Pentium II optimised BLAS library was used for these tests.

It is not easy to see from the above graph but the 128 Kb cache option, as expected, performs on average approximately 1.2 % more efficiently than the default 512 Kb option. From this it was concluded that the best option was to set Mvect=cachesize:131072 (128 Kb) in the *Gaussian 98* make file. All further calculations using the Intel Dual Celeron machine were made using *Gaussian 98* compiled for a 128 Kb level 2 cache.

The results for the different BLAS libraries are given in figure A.2 below. From these results it can be seen that the generic BLAS library provided with *Gaussian 98* performs very poorly with dual processor jobs running almost as slowly as single processor jobs using the Sandia or Portland BLAS libraries. In conclusion it is the Portland BLAS library that performs best averaging approximately 2.3 % higher efficiency over the Sandia libraries for single processor jobs and 3.1 % for dual processor jobs. Based on these results it was therefore decided that the best option for *Gaussian 98* on the IA-32 architecture is to use the Portland Group optimised BLAS library with the data structures aligned for a 128 Kb level 2 cache. All further calculations conducted, both in this benchmark project and in calculations run as part of the myoglobin study, were therefore performed using *Gaussian 98* compiled with these options.

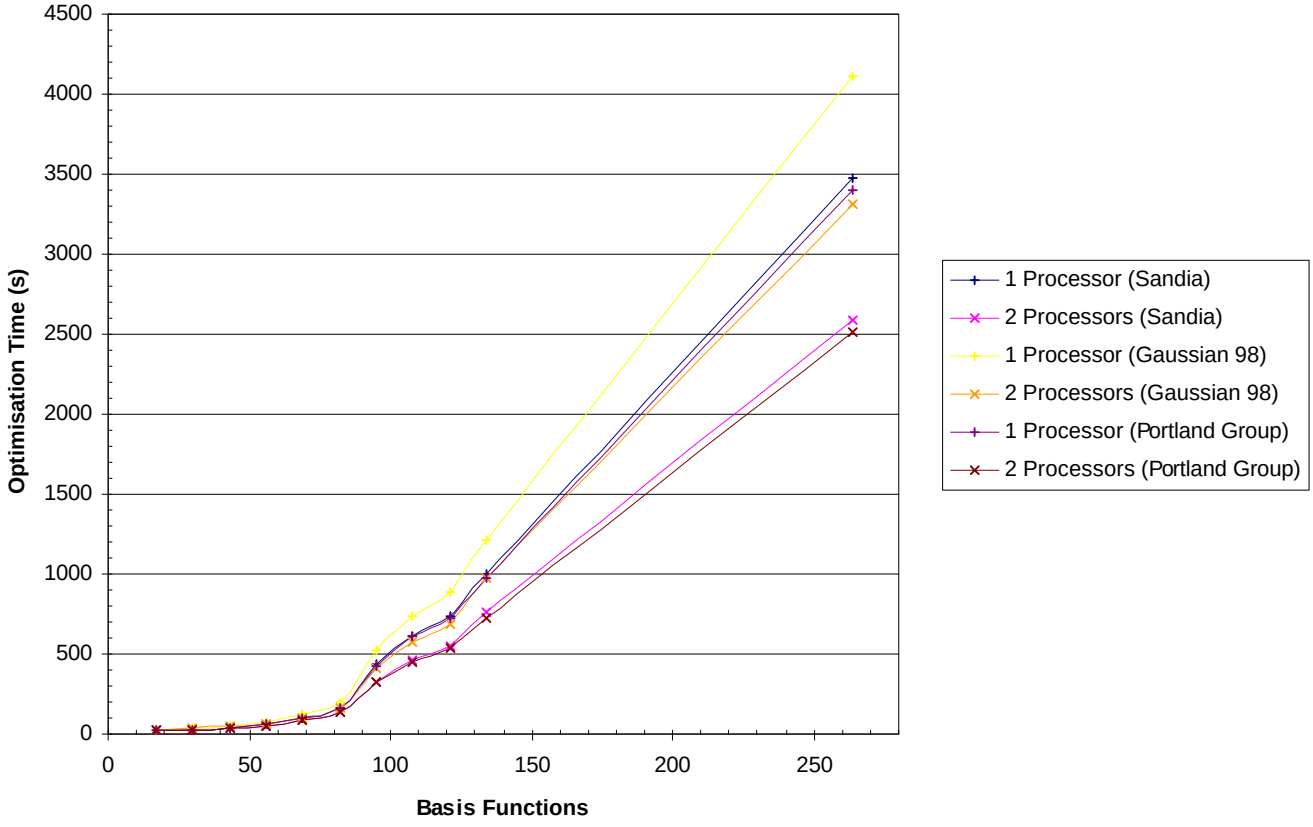


Figure A.2: Plot showing time for geometry optimisation at HF/3-21G as a function of the number of basis functions for a range of straight chain alkanes. The 6 lines correspond to 1 and 2 celeron processors running *Gaussian 98 A.7* compiled with either the default, Sandia or Portland BLAS libraries.

A.3.2 Performance Comparison

A performance comparison was made between single processor and dual processor operation of the Silicon Graphics ONYX 2 and the Dual Intel Celeron machines mentioned in sections *A.1* and *A.2* above. The comparison was made by timing geometry optimisations, using HF/3-21G, of a series of straight chain alkanes running from methane (17 basis functions) to icosane ($C_{20}H_{42}$, 262 basis functions).

Calculations were run using both single processor and dual processor operating modes.

A.3.2.1 Procedure

The computational chemistry package *Gaussian 98 A.7* was used for this investigation. For calculations run on the Silicon Graphics ONYX 2 the *MIPS Pro F77 v7.2* compilers, with the default options, were used to compile *Gaussian 98*. For the calculations run on the Intel Dual Celeron machine it was decided, based on the results described in section *A.3.1.4*, to compile *Gaussian 98* using the optimised parameters described above.

All alkane geometries were drawn in *CambridgeSoft Chem 3D v4.0*, initially optimised using the Allinger MM2 molecular mechanics force field and then exported as Cartesian coordinates. All jobs were allocated 6 MW (48 MB) of ram and the convergence criteria was left at the default of 10^{-8}. The reported CPU time was extracted as reported previously.

A.3.2.2 Results and Discussion

The results obtained are given in figure A.3 on the following page. It can be seen from these results that for large jobs, greater than 100 basis functions, the Silicon Graphics ONYX 2 performs significantly better than the Dual Celeron. Essentially the Dual Celeron running in parallel (2 processors) is roughly equivalent to a single processor of the Silicon Graphics ONYX 2. More insight into the results can be gained by normalising the times with respect to the time taken on the ONYX 2 using equation *A.1*.

$$\text{Efficiency} = \frac{\text{Time on ONYX 2}}{\text{Time on Dual Celeron}} \qquad A.1$$

Plotting these normalised results (*fig. A.4*) shows the performance of the Dual Celeron machine relative to the ONYX 2.

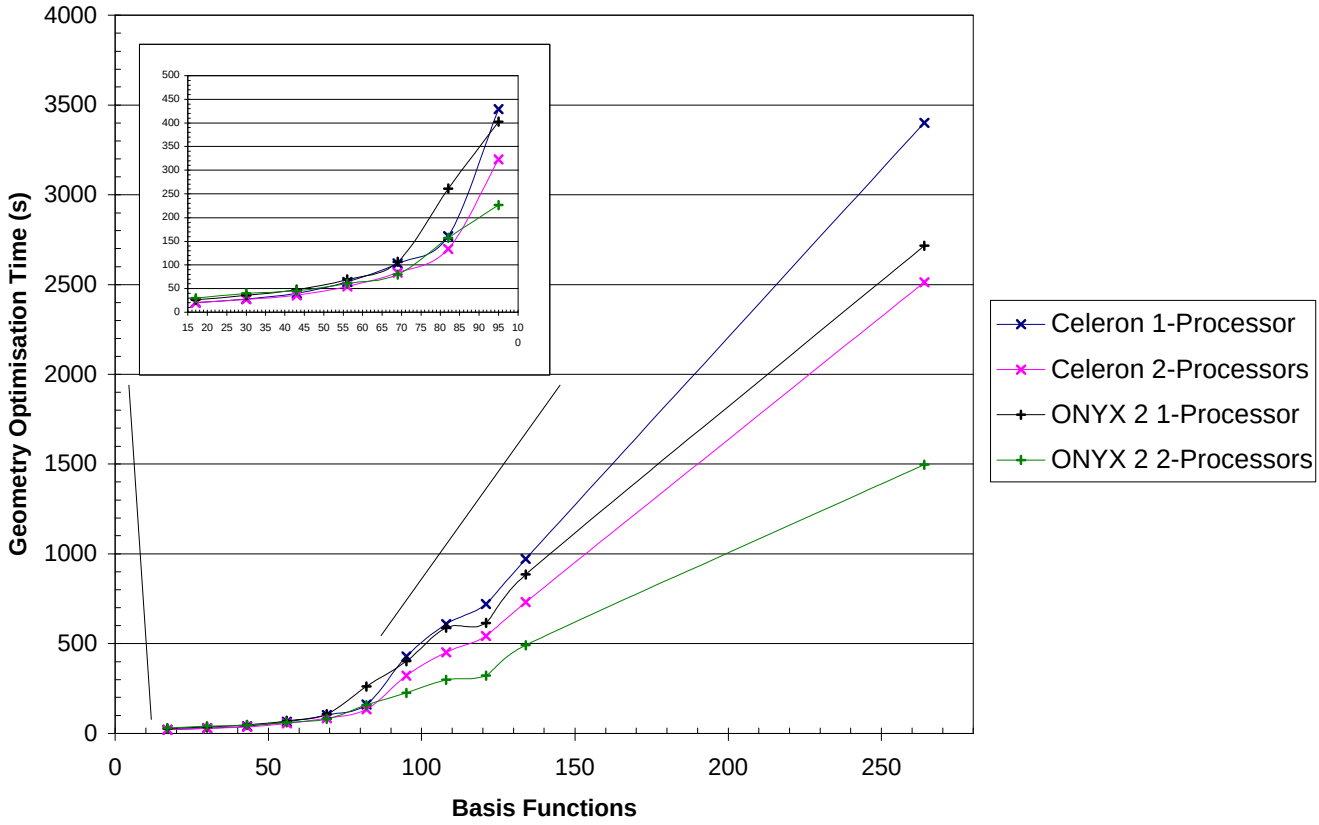


Figure A.3: Plot showing time for geometry optimisation at HF/3-21G as a function of the number of basis functions for a range of straight chain alkanes. The 4 lines correspond to 1 and 2 celeron processors or 1 and 2 processors of a Silicon Graphics ONYX 2.

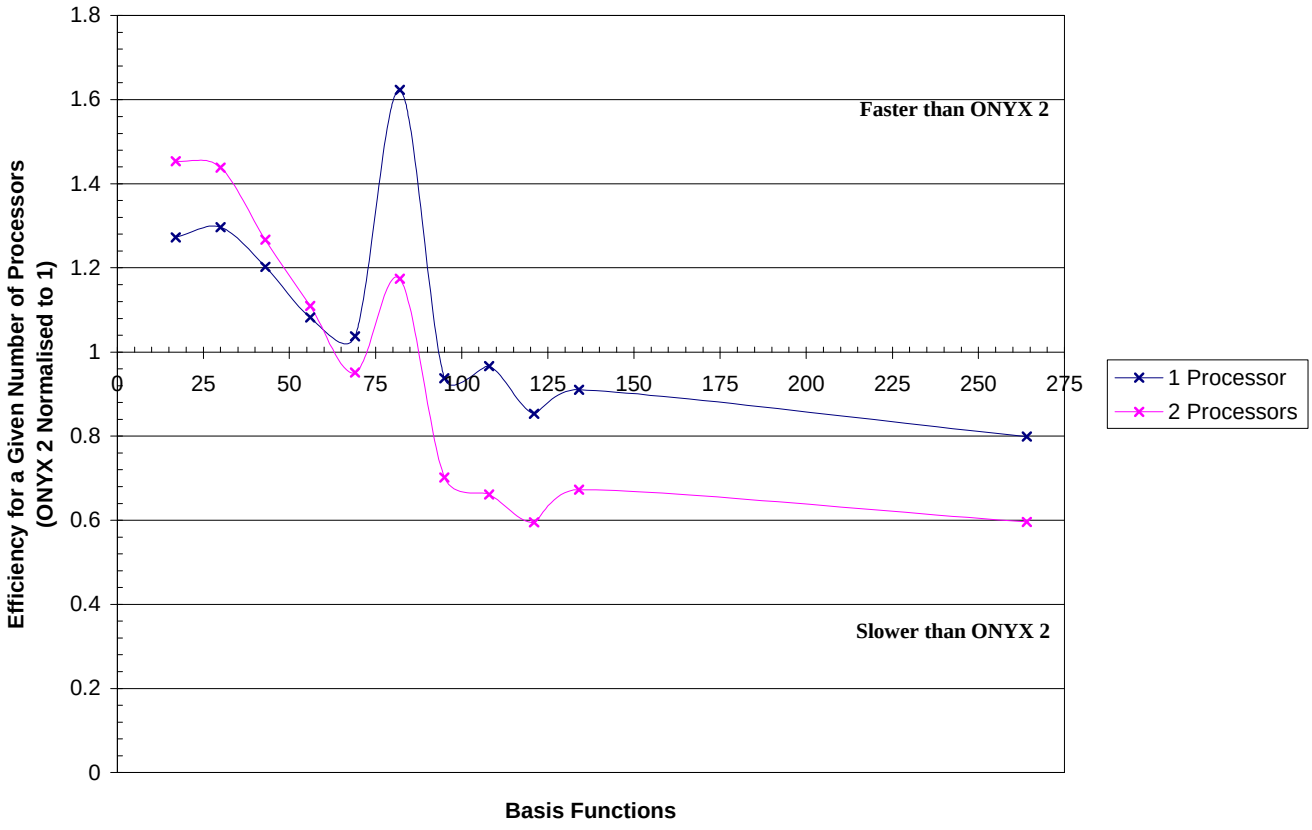


Figure A.4: Plot of efficiency of Celeron machine as a function of the time taken on N Processors of the Silicon Graphics ONYX 2 for geometry optimisation of a range of straight chain alkanes using HF/3-21G.

From figure *A.4* it can be seen that for large systems the ONYX 2 is considerably more efficient and the tendency is for it to become more efficient as the system size grows. It is also clear that the multiprocessing is much more efficient on the ONYX 2 than it is on the Dual Celeron machine. This is due to architectural differences between the two machines. The memory bandwidth, storage I/O speeds, processor interlink speeds and processor cache sizes are all much larger on the Silicon Graphics machine. These improvements manifest themselves in much more efficient parallel processing.

For small systems, however, (< 100 basis functions) the Celeron machine is noticeably more efficient. Why this is the case is not immediately obvious but it most probably stems from the fact that, with a memory allocation of 48 Mb, jobs of less than 100 basis functions will be run in core memory rather than direct. When running in core the jobs are more influenced by the speed at which memory can be read from and written to rather than the speed of the processors. Thus as there are no competing processes running on the Dual Celeron machine, as opposed to the multi user environment on the ONYX 2, both processors have all the system ram to themselves. On the ONYX 2 there is continual congestion on the machine's backbone which could lead to the less than optimum performance. One way to test if this is indeed the case would have been to run the ONYX 2 jobs with the machine restricted to a single user. This option was, however, not available during the course of this research. The large peak at 82 basis functions is largely false in that it represents the transition point for processing via direct methods rather than in core. It would appear that Gaussian 98 has a slightly higher memory overhead on the ONYX 2 that has forced it to proceed via direct methods before the Dual Celeron machine. Thus for 82 basis functions the Dual Celeron machine has run in core while the ONYX 2 has run direct.

In conclusion therefore while the Dual Celeron machine is only one hundredth the cost, per processor, of the ONYX 2 it does not perform as well in parallel. Also the absolute limit for the Celeron configuration is 2 processors while the ONYX 2 can easily be scaled to greater than 128 processors. Having said this the Dual Celeron machine is a very cheap solution and performs reasonably well. It is possible that a cluster of these machines, with efficient load balancing code, could perform exceptionally well for a small price.

Appendix B - Custom Software Developed For This Research

In the process of conducting this research it became apparent that a number of tasks needed to be carried out on a very repetitive basis. One example of this was the extraction of optimised geometries from the *Gaussian 98* output files. For visualisation of the optimised geometries it was necessary to convert these to Brookhaven Protein Databank[B1] (pdb) files. It was also necessary to extract the geometries for performing the BSSE corrections detailed in section *2.9.3* of this report. A number of other repetitive tasks also had to be carried out.

Various applications exist for converting output formats, the most notable being *Babel*[B2] that is available in both Unix and Windows versions. This program is no longer under development but is available free of charge from various chemistry orientated ftp sites. It can inter-convert a huge range of file formats, however, the precise format used in the output was not always the same as that desired in this project. *Babel* also fails to work under *Microsoft Windows 2000* and is unlikely to be updated so that it does. It was therefore decided to develop a computational chemistry tools package specifically for use in Windows that would offer the precise functionality desired for this research project and similar projects being carried out within the research group.

B.1 Computational Chemistry Tools V0.6[B3]

The software detailed below is currently under development with modules being added as the need arises. The version discussed here is, at the time of writing (April 2000), the latest version. The program has been written mainly using *Microsoft Visual Basic 6* with some of the search routines, in the interests of execution speed, written using *Microsoft Visual C++ 6*. The source code is not reproduced here but is available, along with the executables, on the attached support CDRom.

The modules currently available, each of which are discussed separately below, are:

1) Output Conversion
2) Text File Character Extraction
3) Global Search and Replace
4) Time Conversion
5) Unit Conversion

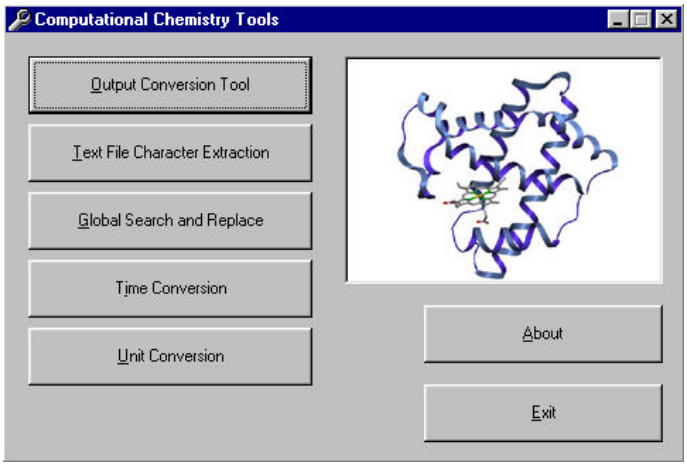


Figure B.1: Computational Chemistry Tools v0.6a2 - Initial startup form showing currently available modules.

B.2 Output Conversion Module

During the course of this research it was continually necessary to extract geometries from *Gaussian* output files. Gaussian Inc. provide a tool for extracting geometries from the checkpoint files produced by *Gaussian 98* (*newzmat*) but this was only available under Unix. The checkpoint files are also very large and so their storage is difficult. For this reason a module was produced for extracting the reported standard orientations in cartesian format from *Gaussian* output files. A screen shot of this module is given in figure B.2, opposite.

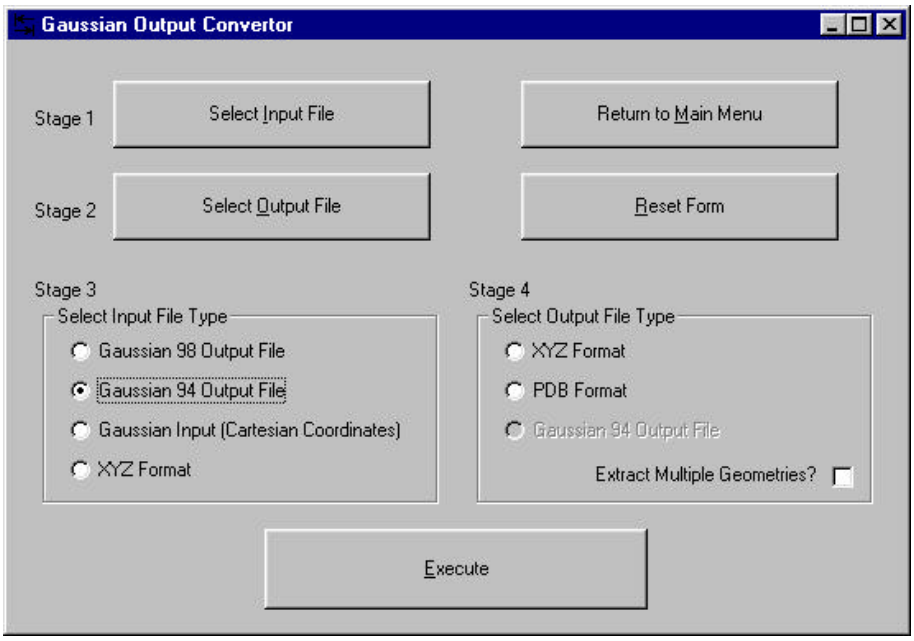


Figure B.2: Computational Chemistry Tools v0.6a2 - Output Conversion module screenshot after an input file and output file have been selected.

The program can currently convert *Gaussian 98* Output, *Gaussian 94* Output and Cartesian *Gaussian* Input files to either PDB or XYZ format. It can also convert from XYZ to PDB and can convert *Gaussian 98* Output files to *Gaussian 94* format. This last task is very useful as it allows *Gaussian 98* files to be easily used with the visualisation tool *Mavis* (which does not support *Gaussian 98*) used in the department.

It is also possible to extract multiple geometries from *Gaussian 98* and *Gaussian 94* files. Selecting this option results in all Standard Orientations being converted to sequentially numbered PDB or XYZ files. One use of this option is to produce a number of PDB files that can then be rendered into a movie showing exactly what atoms were moved as the geometry optimisation progressed. Several examples of such movies are included on the support CDRom.

B.3 Text File Character Extraction

On a number of occasions it was necessary to extract columns of coordinates from a text file. Unix provides an option for doing this in the form of *Awk*, a program included in most Unix distributions, but no such tool exists in *Microsoft Windows*. This module was therefore written to allow up to 5 character ranges to be specified. The program will then extract all characters within these ranges and output them to the selected output file. Such a tool is very useful for extracting

the geometries, without the connection data, from the cartesian coordinate files produced by *CambridgeSoft Chem 3D*. A screenshot of such an extraction is shown below.

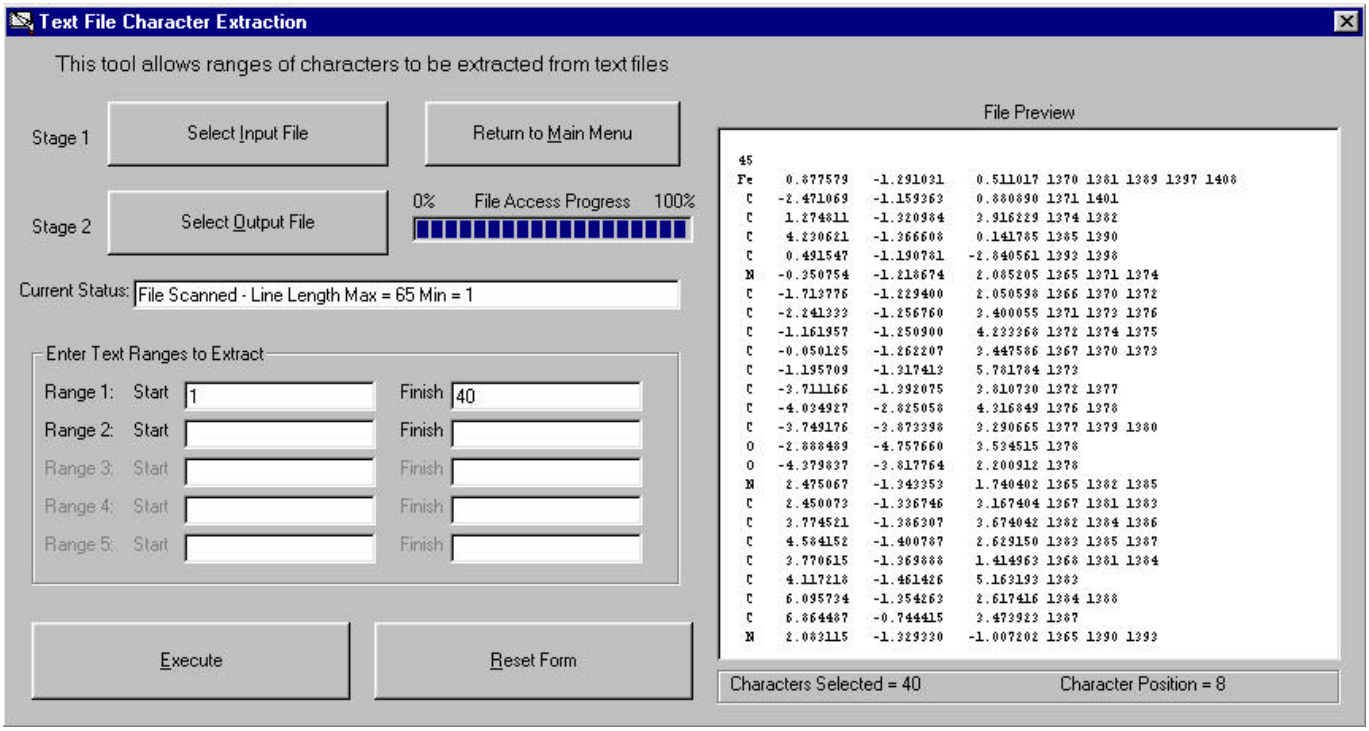


Figure B.3: Computational Chemistry Tools v0.6a2 - Text File Character Extraction module screenshot showing an extraction of characters 1-40 of a Cartesian Coordinate file from Chem3D.

B.4 Global Search and Replace

Most applications in Windows that are designed to handle text offer a search and replace function. However, this generally only applies to the open file. Most programs don't offer the ability to carry out the search and replace over multiple files and if they do it works for only the open files. This means that if a single line has to be replaced in 100 files all the files have to be open at the same time. In the vast majority of cases the system resources are not large enough to permit this.

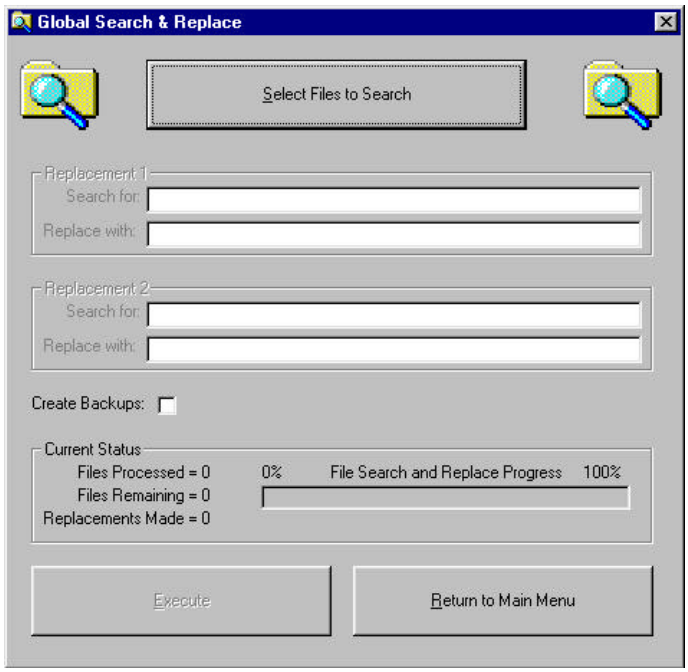


Figure B.4: Computational Chemistry Tools v0.6a2 - Global Search and Replace module prior to the selection of files to search.

The Global Search and Replace module was therefore written to address this problem. A screenshot is shown in figure *B.4* opposite. This module gets around the problem of having every file open by opening each in

turn, making the replacement, and then closing it before opening the next. The program makes use of a temporary file when doing the replacement rather than working on the original files. A power cut, or system crash will therefore not result in data loss. The user can make up to two string replacements at a time and has the option to create backups of every file modified.

The performance of this module was illustrated by a problem that recently affected another member of the research group. A molecular dynamics simulation had been set-up involving 7000 input files. Prior to the running of the simulation it was found that a mistake had been made on one of the lines in every input file. It was therefore necessary to replace this line in all 7000 files or to regenerate all the files from scratch. Using this search and replace module it was possible to correct all 7000 files in less than 25 minutes. Regeneration of all the files from scratch would have taken several hours.

B.5 Time Conversion Module

The number of benchmarking calculations carried out in this research project meant that a convenient method was required to extract the timings reported at the end of each *Gaussian* job to an easily accessible format. The time conversion module, pictured opposite (*fig. B.5*), allows the reported timings to be extracted from a large number of *Gaussian 98* or *Gaussian 94* output files in the *Days Hours Minutes Seconds* format, converted to single unit, plotable, timings (e.g. *Seconds*) and then exported to a single text file. The resulting text file can then easily be imported into a data presentation program such as *Microsoft Excel*.

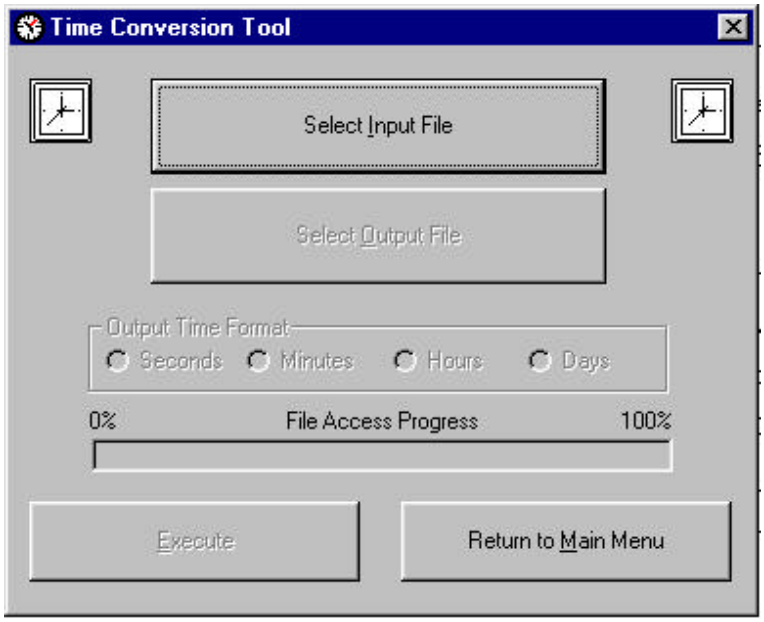


Figure B.5: Computational Chemistry Tools v0.6a2 - Time Conversion module prior to the selection of input files.

In terms of performance this program can easily convert over 500 x 50 Kb long files in less than a minute. This equates to a processing speed of approximately 25 Mb per minute.

B.6 Unit Conversion

This module was written to allow conversion between a range of commonly encountered units in quantum chemistry. For example, *Gaussian 98* reports all its energies in hartrees while most quantum chemistry books and papers report energies in kCal mol^{-1} yet the S.I. units of energy are kJ mol^{-1}. The user selects the input unit and the relevant output unit options (e.g. kJ mol^{-1}, kCal mol^{-1} and eV for hartrees) are enabled. The user then selects the desired output unit and enters the value to be converted in the input box. The output box is automatically updated with the converted value as the user types. A screen shot is shown in figure *B.6* below.

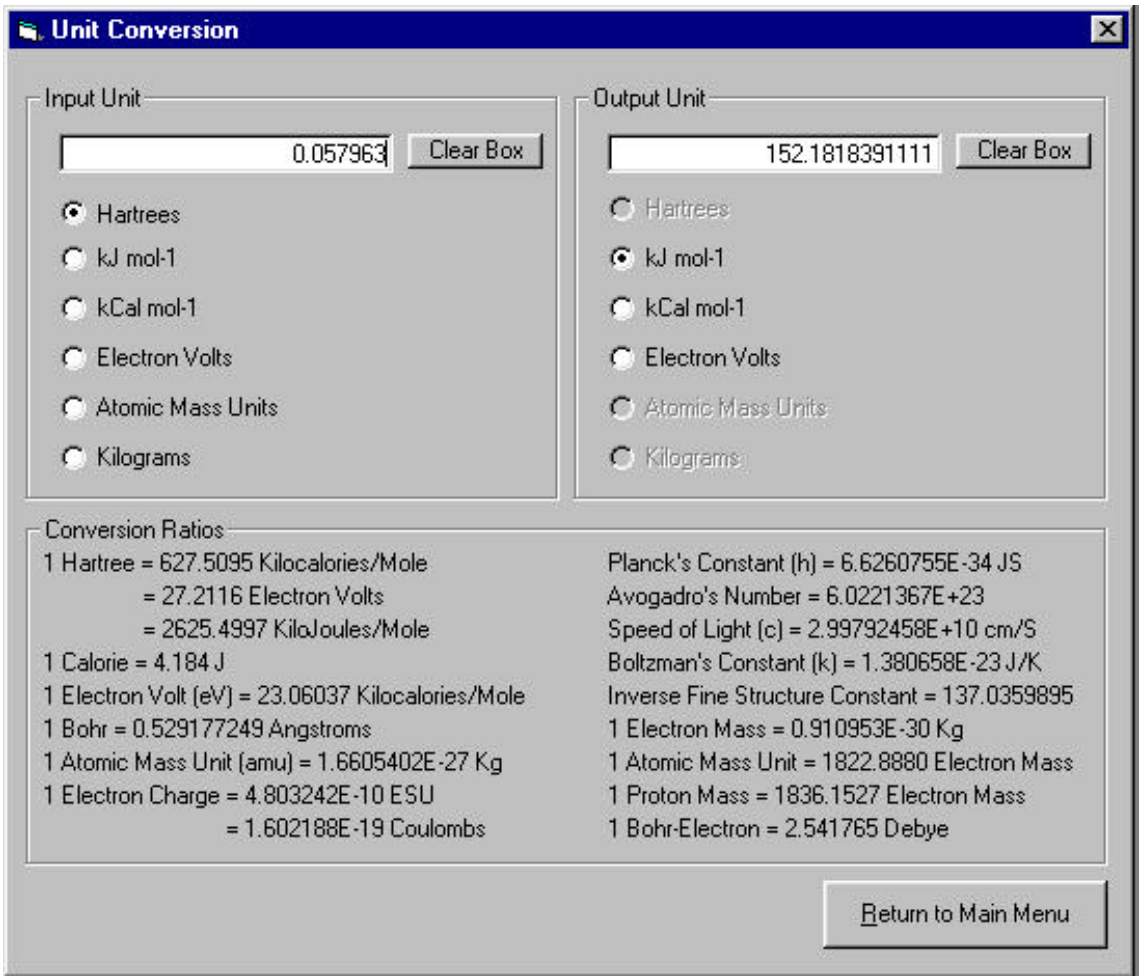


Figure B.6: Computational Chemistry Tools v0.6a2 - Unit Conversion module.

Appendix C – Modification to Gaussian Code (ONIOM)[f]

One of the original aims of this research was to investigate the reactivity of myoglobin using a QM/MM methodology. The original plan was to use the ONIOM method of Morokuma *et al.*[C1] implemented in *Gaussian 98*. The plan was to start by using just the heme section as the quantum mechanical fragment since the heme is not covalently bound to the protein matrix. The size of the QM fragment could then be slowly increased to see how the results changed. In this way it would have been possible to look at how the protein matrix affects ligand specificity and also to predict whether a full *ab initio* calculation would be required to obtain accurate results.

Unfortunately convergence problems, coupled with lack of time and system resources, meant that the QM/MM study had to be shelved. Prior to this decision being taken, however, a number of test calculations highlighted a severe problem with the way in which the ONIOM routines are implemented into *Gaussian 98*.

A number of initial tests showed that a QM/MM calculation using B3P86/3-21G* for the heme unit and the molecular mechanics based universal force field (UFF)[C2] for the rest of myoglobin would require more than 3 GB of ram in order to run. Such high memory demands place severe constraints on the size and speed of calculations. The biggest constraint is that a calculation requiring more than 1024 MB of shared memory cannot be run in parallel with the standard IRIX 6.5 kernel due to the limitations defined by the system constant SHMMAX (0x40000000 [1024 MB]).

Only being able to run the calculations in serial posed a huge problem in that the study would require an unacceptable amount of time to complete. Thus an in-depth study was made of the memory usage within *Gaussian 98* in an attempt to reduce the memory overhead to less than 1024 MB.

C.1 Gaussian Source Code Modification[g]

By studying the *Gaussian 98* source code it was found that the large memory demands were actually a factor of poor programming rather than real memory requirements. For geometry optimisations one of the links called during execution is the Berny optimiser (link 103)[C3]. This link is responsible for determining whether the optimisation has completed or whether another cycle is

[f] The background theory for this section can be found on the support CDRom under *Reports (PDF) -> Appendix C Background.*

[g] Only very small fragments of the source code are reproduced here due to copyright limitations. For full details of the source code please contact Gaussian Inc. (http://www.gaussian.com).

required. Careful debugging of the calculations isolated the memory problems to this link. Line 501 and line 630 (*Gaussian 98 Rev A.7*) of link103.F call the TstCor routine. This routine is responsible for checking that there is sufficient memory available to complete the optimisation. It is not responsible for allocating that memory. Studying the source code highlighted a bug with this memory checking. When using the ONIOM implementation link 103 does not realise that only a small fragment of the molecule is to be treated with quantum mechanics and assumes that the entire molecule (2786 atoms for myoglobin) will be used. This means that it vastly over estimates the amount of memory required.

It was therefore decided to modify these lines and experiment with the memory allocation to find the minimum amount of memory required to run an ONIOM calculation on myoglobin. The original lines from link 103 are as follows:

```
endIf
Call TstCor(IEnd+3*NAtNew,MDV,'Optmz1')
N = NvarRd

endIf
Call TstCor(IEnd,MDV,'Optmz4')
NVar = NvarRd
```

It was found that removing these lines completely led to compilation problems since the results generated by this routine are referenced later in the link. Thus it was decided to modify the lines as follows so that they underestimate the amount of memory required:

```
endIf
Call TstCor((IEnd+3*NatNew)/10,MDV,'Optmz1')
N = NvarRd

endIf
Call TstCor(Iend/10,MDV,'Optmz4')
NVar = NvarRd
```

This minor modification to the source code resulted in the minimum memory required to run the ONIOM job being reduced from over 3GB to approximately 780 MB. This reduction of over 75 % meant that calculations could be run in parallel without modification of the machine's kernel.

The only problem with these modifications is that where previously allocating too little memory to a job would result in an allocation failure and graceful exit of *Gaussian 98*, now allocating too little memory results in a segmentation fault and thus corruption of the checkpoint file. In most situations, however, this would not pose a problem as any such segmentation fault would occur within the first 30 seconds or so of starting a job and so would lead to little data loss.

With these modifications a QM/MM study of myoglobin, assuming the convergence problems could be solved, using the ONIOM formalism is now feasible and could form the basis of any future work.

Appendix D – Project Support CDRom

A large amount of numerical data was produced in the process of carrying out this research. Most of this material is not immediately relevant to understanding the conclusions made within this project but it is important to someone who wants to look at the results in more detail and possibly adapt some of the work done in this project.

It was therefore decided to include all of the data generated in this project on a support CDRom to be distributed with the hard copy of the report. Merely including the data in its raw form on the CD was not a viable solution as navigation to someone unfamiliar with the layout of the CDRom would prove impossible. Thus it was decided to create a web front end that would provide easy navigation of all the material included on the CDRom.

The CDRom is accessed by selecting index.htm from the root directory of the CDRom drive. This presents an initial introduction screen. Continuing from this screen the user is presented with an expandable/collapsible menu as illustrated in figure D.1 below.

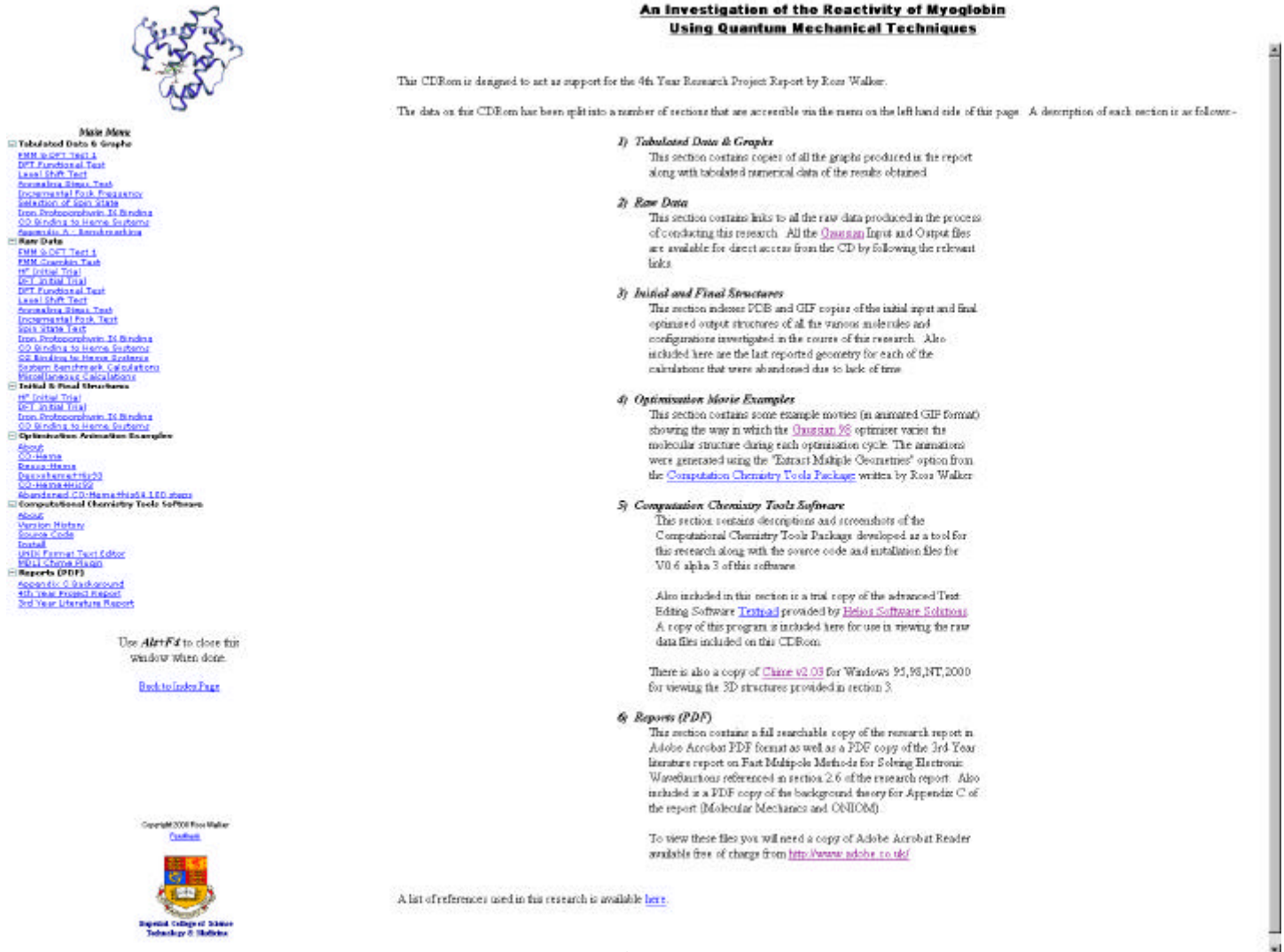


Figure D.1: Screenshot showing the navigation menu and introduction screen for the project support CDRom.

The available sections on the CDRom are as follows:

1) Tabulated Data & Graphs

This section contains copies of all the graphs produced in the report along with tabulated numerical data of the results obtained.

2) Raw Data

This section contains links to all the raw data produced in the process of conducting this research. All the Gaussian Input and Output files are available for direct access from the CD by following the relevant links.

3) Initial and Final Structures

This section indexes PDB and GIF copies of the initial input and final optimised output structures of all the various molecules and configurations investigated in the course of this research. Also included here are the last reported geometry for each of the calculations that were abandoned due to lack of time.

4) Optimisation Movie Examples

This section contains some example movies (in animated GIF format) showing the way in which the Gaussian 98 optimiser varies the molecular structure during each optimisation cycle. The animations were generated using the "Extract Multiple Geometries" option from the Computation Chemistry Tools Package written by Ross Walker.

5) Computation Chemistry Tools Software

This section contains descriptions and screenshots of the Computational Chemistry Tools Package developed as a tool for this research along with the source code and installation files for V0.6 alpha 3 of this software.

Also included in this section is a trial copy of the advanced Text Editing Software Textpad provided by Helios Software Solutions. A copy of this program is included here for use in viewing the raw data files included on this CDRom.

There is also a copy of Chime v2.03 for Windows 95,98,NT,2000 for viewing the 3D structures provided in section 3.

6) Reports (PDF)

This section contains a full searchable copy of the research report in Adobe Acrobat PDF format as well as a PDF copy of the 3rd Year literature report on Fast Multipole Methods for Solving Electronic Wavefunctions referenced in section 2.6 of the research report. Also included is a PDF copy of the background theory for Appendix C of the report (Molecular Mechanics and ONIOM).

To view these files you will need a copy of Adobe Acrobat Reader available free of charge from http://www.adobe.co.uk/

Please note: The CDRom web pages are best viewed full screen at a resolution of 1024x768 or above and in 16 bit colour or above.

References

1. Adapted from Maurus, R. *et al. Biophys. Bioch. Acta.* 1341, (1997)

2. Cited in - Royal Swedish Academy of Sciences - 1998 Nobel Prize Press Release.

3. T. E. Creighton, "Proteins - Structures and Molecular Properties", Freeman, **1984** p. 1.

4. *Ibid.* pp. 6-7.

5. G. Hill, J. Holman, "Chemistry in Context", Edn. 3, Nelson, **1989**, p. 364.

6. G. P. Young, P. Murthi, M. A. Levitt, Y. Gawad, "Serial use of bedside CKMB/myoglobin device to detect acute myocardial infarction in emergency department chest pain patients", *J. Emerg. Med.*, 17, **1999**, pp. 769-775.; M. Plebani, M. Zaninotto, "Cardiac markers: present and future", *Int. J. Clin. & Lab. Res.*, 29, **1999**, pp. 56-63.; R. Valdes, S. A. Jortani, "Standardizing utilization of biomarkers in diagnosis and management of acute cardiac syndromes", *Clinica Chimica Acta*, 184, **1999**, pp. 135-140.

7. Kendrew, J. in Watson, H.C., "The Stereochemistry of the Protein Myoglobin", *Prog. Stereochem.*, **4**, 1969, 299.

8. Collman, J.P., Brauman, J.I., Iveson, B.L., Sessler, J.L., Morris, R.M., Gibson, Q.H., *J. Am. Chem. Soc.*, **105**, 1983, 3052-3064.; Traylor, T.G., Koga, N., Deardurff, L.A., Swepston, P.N., Ibers, J.A., *J. Am. Chem. Soc.*, **106**, 1984, 5132-5143.; Springer, B.A., Egeberg, K.D., Silgar, S.G., Rohlfs, R.J., Matthews, A.J., Olson, J.S., *J. Biol. Chem.*, **264**, 1989, 3057-3060.

9. Kuriyan, J., Wilz, S., Karplus, M., Petsko, G.A., *J. Mol. Biol.*, **192**, 1986, 133-154.

10. *Ibid.*

11. Collman, J.P., Brauman, J.I., Halbert, T.R., Suslick, K.S., *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 1976, 3333-3337.

12. Case, D.A., Karplus, M., *J. Mol. Biol.*, **123**, 1978, 697-701.

13. *Ibid.*

14. Kuriyan, J., Wilz, S., Karplus, M., Petsko, G.A., *J. Mol. Biol.*, **192**, 1986, 133-154.

15. Huber, R., Epp, O., Formanek, H., *J. Mol. Biol.*, **52**, 1970, 349-354.

16. Norvell, J.C., Nunes, A.C., Schoenborn, B.P., *Science*, **190**, 1975, 568-569

17. Quillin, M.L., Arduini, R.M., Olson, J.S., Phillips Jr., G.N., *J. Mol. Biol.*, **234**, 1993, 140-155.

18. Springer, B.A., Sligar, S.G., Olson, J.S., Phillips Jr., G.N., *Chem. Rev.*, **94**, 1994, 699-714.

19. Jewsbury, P., Yamamoto, S., Minato, T., Saito, M., Kitagawa, T., *J. Am. Chem. Soc.*, **116**, 1994, 11586-11587.

20. Pauling, L., *Nature*, **203**, 1964, 182-183.

21. Phillips, S.E., Schoenborn, B.P., *Nature*, **292**, 1981, 81-82.

22. Hanson, J.C., Schoenborn, B.P., *J. Mol. Biol.*, **153**, 1981, 117-146.

23. Schrödinger, E. *Ann. Phys.*, **29**, 361 (1926)

24. Atkins, P.W., Friedman, R.S., "Molecular Quantum Mechanics", Oxford, Edn.3, 1997, P. 17.

25. Born, M. & Oppenheimer, R. *Ann. Phys.* **84**, 457 (1927)

26. *Ibid.*

27. Hartree, D. *Proc. Cam. Phil. Soc.* **24**, 89 (1928)

28. Hehre, W.J., Radom, L., Schleyer, P. & Pople, J.A., "*Ab Initio* Molecular Orbital Theory", John Wiley & Sons, Inc., 1986.

29. Atkins, P.W., Friedman, R.S., "Molecular Quantum Mechanics", Oxford, Edn.3, 1997, P. 219.

30. Szabo, A. & Ostlund, N.S., "Modern Quantum Chemistry", McGraw-Hill, Inc., 1989.

31. *Ibid.*

32. Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.

33. Roothaan, C.C.J., *Rev. Mod. Phys.*, **23**, 1951, 69

34. Hall, G.G., *Proc. R. Soc. (London)*, **A205**, 1951, 541

35. Slater, J.C., *Phys. Rev.*, **36**, 1930, 57.

36. Boys, S.F., *Proc. R. Soc. (London)*, **A200**, 1950, 542

37. Hehre, W.J., Stewart, R.F., Pople, J.A., *J. Chem. Phys.*, **51**, 1969, 2657.

38. Szabo, A. & Ostlund, N.S., "Modern Quantum Chemistry", McGraw-Hill, Inc., 1989, P. 156.

39. Dunning Jr., T.H., Hay, P.J. "Modern Theoretical Chemistry", Ed. H.F. Schaefer, III, Plenum, New York, 1996, vol. 3, p.1.

40. Binkley, J.S., Pople, J.A., Hehre, W.J., *J. Am. Chem. Soc.*, **102**, 1980, 939.

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

42. Krishnan, R., Binkley, J.S., Seeger, R., Pople, J.A., *J. Chem. Phys.* **72**, 1980, 650.

43. Pietro, W.J., Francl, M.M., Hehre, W.J., DeFrees, D., Pople, J.A., Binkley, J.S., *J. Am. Chem. Soc.*, **104**, 1982, 5039.

44. Hariharan, P.C., Pople, J.A., *Theor. Chim. Acta.*, **28**, 1973, 213.

45. *Ibid.*

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

47. Stone, A.J., "The Theory of Intermolecular Forces", Clarendon Press, 1996.

48. Foresman, J.B., Frisch, Æ., "Exploring Chemistry with Electronic Structure Methods", Gaussian Inc., Edn. 2, 1996.

49. Shavitt, I., *in* Schaefer, H.F., *ed.*, "Methods of Electronic Structure Theory", Plenum Press., 1977.

50. Hehre, W.J., Radom, L., Schleyer, P. & Pople, J.A., "*Ab Initio* Molecular Orbital Theory", John Wiley & Sons, Inc., 1986.

51. Hohenburg, P., Kohn, W., *Phys. Rev.*, **136**, 1964, B864.

52. Parr, R.G., Yang, W. "Density Functional Theory", OUP, 1989; Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.

53. Block, F., *Z. Physik*, **57**, 1929, 545, *in* [Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.]

54. Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.

55. Kohn, W., Sham, L.J., *Phys. Rev.*, **140**, 1965, A1133.

56. Becke, A.D., *Phys. Rev.*, **A38**, 1988, 3098.

57. Lee, C., Yang, W., Parr, R.G., *Phys. Rev.*, **B37**, 785, 1988.

58. Johnson, B.G., Frisch, M.J., *J. Chem. Phys.*, **100**, 1994, 8448.

59. Challacombe, M. & Schwegler, E. *J.Chem.Phys.*, **106**, 13, 1997, pp. 5526-5536

60. Gaussian 98, Revision A.7, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian, Inc., Pittsburgh PA, **1998**.

61. van Duijneveldt, F.B., van Duijneveldt-van de Rijdt, J.G.C.M., van Lenthe, J.H., *Chem. Rev.*, **94**, 1994, 1873.

62. Frisch, Æ., Frisch,M.J., "Gaussian 98 User's Reference", Edn 6, Gaussian Inc. **1998**, p. 3.

63. Hehre, W.J., Ditchfield, R., Pople, J.A., *J. Chem. Phys.*, **56**, 1972, p. 2257.

64. Becke, A.D., *Phys. Rev.*, **A38**, 1988, 3098.

65. Lee, C., Yang, W., Parr, R.G., *Phys. Rev.*, **B37**, 785, 1988.

66. Millam, J.M., Scuseria, G.E., *J. Chem. Phys.*, **106**, 1997, 5569; Burant, J.C., Scuseria, G.E., Frisch, M.J., *J. Chem. Phys.*, **195**, 1996, 8969; Burant, J.C., Strain, M. C., Scuseria, G.E., Frisch, M.J., *CPL*, **258**, 1996, 45; Burant, J.C., Strain, M.C., Scuseria, G.E., Frisch, M.J., *Chem. Phys. Lett.*, **248**, 1996, 43; Strain, M.C., Scuseria, G.E., Frisch, M.J., *Science*, **271**, 1996, 51.

67. Allinger, N.L., *J. Am. Chem. Soc.*, **99**, 1977, 8127.

68. Maurus, R., Overall, C.M., Bogumil, R., Luo, Y., Mauk, A.G., Smith, M., Brayer, G.D., *Biochim. Biophys. Acta.*, **1341**, 1997, pp. 1.

69. Lide, D.R., "CRC Handbook of Chemistry and Physics", Boca Raton London: CRC Press, **1996**.

70. Becke, A.D., *J. Chem. Phys.*, **98**, 1993, 5648.

71. Lee, C., Yang, W., Parr, R.G., *Phys. Rev. B*, **37**, 1988, 785.

72. Schafer, A., Hann, H., Ahlrichs, R., *J. Chem. Phys.*, **97**, 1992, 2571.

73. Giffiths, E., PhD. Thesis, Imperial College of Science Technology and Medicine, London, **2000**.

74. Perdew, J.P., *Phys. Rev. B*, **33**, 1986, 8822.

75. *Ibid*.

76. Cheng, X., Schoenborn, B.P., *Acta. Crystallogr. Sect. B.*, **46**, 1990, 195.

77. Burke, K., Perdew, J.P., Wang, Y., in "Electron Density Functional Theory: Recent Progress and New Directions", Ed. Dobson, J.F., Vignale, G., Das, M.P., Plenum, **1998**.

78. Perdew, J.P., in "Electronic Structure of solids '91", Ed. P. Ziesche, Eschrig, H., Akademie Verlag, Berlin, **1991**, p. 11.

79. Perdew, J.P., Chevary, J.A., Vosko, S.H., Jackson, K.A., Pederson, M.R., Singh, D.J., Fiolhais, C., *Phys. Rev. B.*, **46**, 1992.

80. Perdew, J.P., Chevary, J.A., Vosko, S.H., Jackson, K.A., Pederson, M.R., Singh, D.J., Fiolhais, C., *Phys. Rev. B.*, **48**, 1993.

81. Perdew, J.P., Burke, K., Wang, Y., *Phys. Rev. B.*, **54**, 1996, 16533.

82. Adamo, C., Barone, V., *J. Comp. Chem.*, **19**, 1998, 419.

83. Gill, P.M.W., *Mol. Phys.*, **89**, 1996, 433.

84. Saunders, V.R., Hillier, I.H., *Mol. Phys.*, **28**, 1974, 819.

85. Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.

86. Kirkpatrick, S., Gelatt Jr, C.D., Vecchi, M.P., *Science*, **220**, 1983, 671; Wilson, S.R., Cui, W., *Biopolymers*, **29**, 1990, 225.

87. Oganesyan, V.S., Sharonov, Y.A., *Spectrochimica Acta Part A.*, **53**, 1997, 433-449.

88. Huynh, B.H., Papaefthymiou, G.C., Yen, C.S., Groves, J.L., Wu, C.S., *J. Chem. Phys.*, **61**, 1974, 3750-3758.

89. Challacombe, M., *Personal Correspondence*. – Unpublished work.

90. Giffiths, E., PhD. Thesis, Imperial College of Science Technology and Medicine, London, **2000**.

91. Lide, D.R., "CRC Handbook of Chemistry and Physics", Boca Raton London: CRC Press, **1996**.

92. Jewsbury, P., Yamamoto, S., Minato, T., Saito, M., Kitagawa, T., *J. Am. Chem. Soc.*, **116**, 1994, 11586-11587.

93. Springer, B.A., Sligar, S.G., Olson, J.S., Phillips Jr., G.N., *Chem. Rev.*, **94**, 1994, 699-714.

94. *Ibid.*

95. Jewsbury, P., Yamamoto, S., Minato, T., Saito, M., Kitagawa, T., *J. Am. Chem. Soc.*, **116**, 1994, 11586-11587.

96. Hay, P.J., Wadt, W.R, *J. Chem. Phys.*, **82**, 19845, 299-310.

97. Jensen, F., "Introduction to Computational Chemistry", John Wiley & Sons, 1999.

A1. Walker, R. *Unpublished Experimental Work.*

A2. US Department of Energy – Accelerated Strategic Computing Initiative (ASCI). Sandia National Labs, ASCI Red Pentium II BLAS 1.2F, http://www.cs.utk.edu/~ghenry/distrib/archive.htm#blas

B1. Royal Society of Chemistry, Brookhaven Protein Databank, http://www.rcsb.org/pdb/

B2. Walters, P., Stahl, M., Babel v1.6, (babel@mercury.aichem.arizona.edu).

B3. Walker, R., "Computational Chemistry Tools v0.6", (r.c.walker@ic.ac.uk), 2000.

C1. Dapprich, S., Komáromi, I., Byun K., Morokuma, K., Frisch, M., *Theo. Chem. J. Mol. Struc.*, **461-462**, 1999, 1-21.; Maseras, F., Morokuma, K., *J. Comp. Chem.*, **16**, 1995, 1170-1179.; Svensson, M., Humbel, S., Froese, R., Matsubara, T., Sieber, S., Morokuma, K., *J. Phys. Chem.*, **100**, 1996, 19357-19363.; Matsubara, T., Masera, F., Koga, N., Morokuma, K., *J. Phys. Chem.*, **100**, 1996, 2573-2580.

C2. Rappé, A.K., Casewit, C.J., Colewell, K.S., Goddard III, W.A., Skiff, W.M., *J. Am. Chem. Soc.*, **114**, 10024 (1992).

C3. Frisch, Æ., Frisch, M.J., "Gaussian 98 User's Reference", Edn 6, Gaussian Inc. **1998**, p. 3.

Acknowledgements:

I would like to thank my supervisor Ian Gould for his help and computers. Julian Gale for his help in Ian Gould's absence. Sam for proof reading, suggestions and not hesitating to lend me her swipe card. Richard for welcome distractions. Ed for his technical help and Halima for her extensive assistance and for providing me with numerous opportunities to hone my computer support skills.