Bridging the gap between computational chemistry and macromolecular crystallography

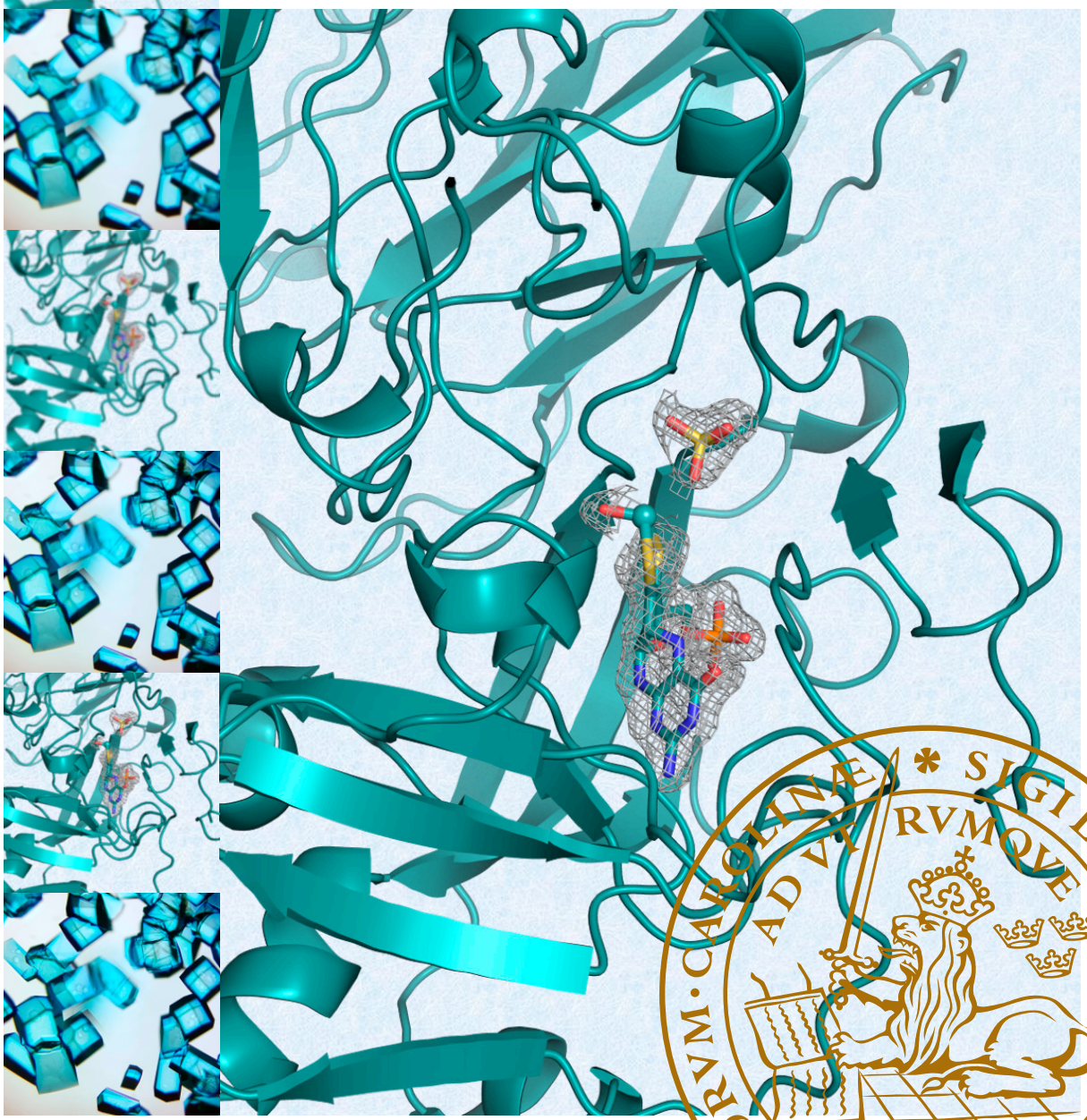
Caldararu, Octav

2019

Document Version:
Publisher's PDF, also known as Version of record

Bridging the gap between computational chemistry & macromolecular crystallography

OCTAV CALDARARU | DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY



Bridging the gap between computational chemistry and macromolecular crystallography

Octav Caldararu

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Science, Lund University, Sweden.
To be defended at
Centre for Chemistry and Chemical Engineering, Hall F
6th December 2019, 9:00

Faculty opponent
Dr. Garib Murshudov, MRC Laboratory of Molecular Biology, Cambridge, UK

Organization LUND UNIVERSITY Centre for Chemistry and Chemical Engineering P.O. Box 124, 221 00, Lund, Sweden	**Document name** **DOCTORAL DISSERTATION** **Date of issue** 2019-12-06
Author(s) Octav Caldararu	Sponsoring organization

Title and subtitle
Bridging the gap between computational chemistry and macromolecular crystallography

Abstract

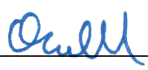
Knowledge of the atomic structure of biomolecules, such as proteins, is paramount to understanding their function and interactions in the human body. For example, knowledge of the atomic structure of a target protein is crucial for developing drugs that bind strongly to it and thus help cure diverse diseases. Macromolecular crystallography is the forefront method for determining the atomic structure of proteins, especially through X-ray diffraction experiments. However, the data obtained from these experiments are not the atomic structure but need to be processed and interpreted before arriving at the individual positions of atoms in a protein. This intepretation is done through computational techniques that share some of the algorithms and problems with computational chemistry.

In this thesis, we use several methods that combine computational chemistry and macromolecular crystallography for the study of multiple important proteins. Crystallographic refinement combined with quantum mechanical calculations (quantum refinement) is used to improve the X-ray structures of three metalloenzymes. Furthermore, a quantum refinement procedure for neutron structures is developed and applied to two important enzymes. We also investigate how to use and improve the existing information on dynamics from crystallography experiments. To this end, we test whether conformational entropy can be calculated directly from B-factors. Additionally, ensemble refinement is used to explore ligand dynamics in the binding site of galectin-3 and reveals hidden conformations that were not apparent in traditional crystallographic refinement methods. Finally, we study the modeling of water molecules in protein X-ray and neutron crystal structures. We show that molecular dynamics simulations can reproduce crystal water molecules, if protein movements are correctly taken into account. Moreover, we have developed a method to automatically improve the orientation of water molecules in neutron structures.

Key words
Protein structure, X-ray crystallography, Neutron crystallography, Quantum mechanics, QM/MM, Quantum refinement, Ensemble refinement, Water structure

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language** English
ISSN and key title		**ISBN** 978-91-7422-702-4 (print)
Recipient's notes	**Number of pages** 238	Price
	Security classification	

Signature  Date 2019-10-25

Bridging the gap between computational chemistry and macromolecular crystallography

Octav Caldararu

Coverphoto by Ada-Ioana Bunea

Faculty of Science
Department of Theoretical Chemistry

ISBN 978-91-7422-702-4 (print)
ISBN 978-91-7422-703-1 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2019

“These atoms are liars”

Dan Barrett

Table of Contents

List of publications

Paper I: Cao, L.; Caldararu, O.; Rosenzweig, A.C.; Ryde, U. Quantum refinement does not support dinuclear copper sites in crystal structures of particulate methane monooxygenase. *Angew. Chem. - Int. Ed.* **2018**, *57*, 162–166.

Paper II: Cao, L.; Caldararu, O.; Ryde, U. Protonation states of homocitrate and nearby residues in nitrogenase studied by computational methods and quantum refinement. *J. Phys. Chem. B* **2017**, *121*, 8242–8262.

Paper III: Caldararu, O.; Feldt, M.; Cioloboc, D.; Van Severen, M.-C.; Starke, K.; Mata, R. A.; Nordlander, E.; Ryde, U. QM/MM Study of the reaction mechanism of sulfite oxidase. *Sci. Rep.* **2018**, *8*, 4684.

Paper IV: Caldararu, O.; Manzoni, F.; Oksanen, E.; Logan, D.T.; Ryde, U. Refinement of protein structures using a combination of quantum-mechanical calculations with neutron and X-ray crystallographic data. *Acta Cryst. D.* **2019**, *75*, 368–380.

Paper V: Caldararu, O.; Oksanen, E.; Ryde, U.; Hedegård, E.D. Mechanism of hydrogen peroxide formation by lytic polysaccharide monooxygenase. *Chem. Sci.* **2019**, *10*, 576–586.

Paper VI: Kelpsas, V.; Caldararu, O. ; Kulkani, Y.; Wierenga, R.; Kamerlin, L.; Ryde, U.; Wachenfeldt C.; Oksanen E. Neutron structures of *Leishmania mexicana* triose phosphate isomerase with reaction-intermediate mimics shed light on the chemical step. *Manuscript.*

Paper VII: Caldararu, O.; Ekberg, V.; Oksanen, E.; Ryde, U. Exploring ligand dynamics in protein crystal structures with ensemble refinement. *Manuscript.*

Paper VIII: Caldararu, O.; Kumar, R; Oksanen, E.; Logan, D.T.; Ryde, U. Are crystallographic B-factors suitable for calculating protein conformational entropy? *Phys. Chem. Chem. Phys.* **2019,** *21*, 18149–18160.

Paper IX: Caldararu, O.; Misini Ignjatović, M; Oksanen, E; Ryde, U. Water structure in solution and crystal molecular dynamics simulations compared to protein crystal structures. *Manuscript.*

Paper X: Eriksson, A.; Caldararu, O.; Oksanen, E.; Ryde, U. Automated orientation of water molecules in protein neutron-diffraction structures. *Manuscript.*

List of publications not included in the thesis

1. Misini Ignjatović, M.; Caldararu, O.; Dong, G.; Muñoz-Gutierrez, C.; Adasme-Carreño, F.; Ryde, U. Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations. *J. Comput. Aided. Mol. Des.* **2016**, *30*, 707-730.
2. Caldararu, O.; Olsson, M.; Riplinger, C.; Neese, F.; Ryde, U. Binding free energies in the SAMPL5 octa-acid host–guest challenge calculated with DFT-D3 and CCSD(T). *J. Comput. Aided. Mol. Des.* **2017**, *31*, 87–106.
3. Caldararu, O.; Olsson, M.A.; Misini Ignjatović, M.; Wang, M.; Ryde, U. Binding free energies in the SAMPL6 octa-acid host–guest challenge calculated with MM and QM methods. *J. Comput. Aided. Mol. Des.* **2018**, *32*, 1027-1046.
4. Cao, L.; Caldararu, O.; Ryde, U. Protonation and reduction of the FeMo cluster in nitrogenase studied by quantum mechanics/molecular mechanics (QM/MM) calculations. *J. Chem. Theory Comput.* **2018**, *14*, 6653–6678.
5. Verteramo, M.L.; Stenström, O.; Misini Ignjatović, M.; Caldararu, O.; Olsson, M.A. Manzoni, F.; Leffler, H.; Oksanen, E.; Logan, D.T.; Nilsson, U.J.; Ryde, U.; Akke, M. Interplay between conformational entropy and solvation entropy in protein–ligand binding. *J. Am. Chem. Soc.* **2019**, *141*, 2012-2026.
6. Cao, L.; Börner, M.C.; Bergmann, J; Caldararu, O; Ryde, U. Geometry and electronic structure of the P-cluster in nitrogenase studied by QM/MM and quantum refinement. *Inorg. Chem.* **2019***, 58*, 9672-9690.

List of article contributions

Paper I: I performed crystallographic evaluations and analysed the quantum-refinement calculations. I participated in writing the manuscript

Paper II: I performed crystallographic evaluations and analysed the quantum-refinement calculations.

Paper III: I performed QM/MM calculations of the oxidised form of the cofactor and all quantum-refinement calculations. I participated in writing the manuscript.

Paper IV: I performed all the final refinements and quantum-refinement calculations. I participated in writing the manuscript.

Paper V: I performed all the quantum-refinement calculations and all the QM/MM calculations. I wrote the first draft of the manuscript.

Paper VI: I performed all the quantum-refinement calculations and all the QM/MM calculations. I participated in writing the manuscript.

Paper VII: I performed all the calculations. I wrote the first draft of the manuscript.

Paper VIII: I performed all molecular dynamics simulations, all refinements and all entropy calculations. I took part in writing the code to calculate entropies. I wrote the first draft of the manuscript.

Paper IX: I took part in designing the project. I performed all molecular dynamics simulations. I wrote the first draft of the manuscript.

Paper X: I designed the project. I wrote the code to optimise water orientations. I supervised the BSc student doing the calculations. I participated in writing the manuscript.

Popular science summary

The understanding of how the human body and other multicellular organisms function is one of the biggest current scientific tasks. The final goal is to find treatment for diseases, slowing down aging and in general providing a better quality of life, but we cannot achieve these goals without a deep knowledge of how the parts function together at a detailed level. All organisms are made up of cells, which in turn are made up of a variety of molecules, which are made up of atoms arranged and bonded together in different ways. Therefore, the most detailed level we can go to in biology is the atomic level.

Proteins are the most versatile of the biological molecules. Each cell contains tens of thousands of different proteins, each performing a certain function. Proteins are composed of a combination of building blocks, called amino acids, which arrange themselves in three dimensions to provide stability, localisation and function. The structure of proteins is a chemical problem, as the amino acids interact through covalent chemical bonds and through non-bonded interactions. It is known that the atomic structure of proteins influences their function. Explaining this structure-function relationship is still challenging for many proteins, as we require better understanding to do so, which is why this topic is a research focus in both biology and chemistry.

Thus, we need to determine the structures of as many important proteins as we can and understand how this contributes to their function on a case-by-case basis. If we do achieve this understanding, we can then try to modify proteins such that they perform that function better or we can design drugs that can bind to specific target proteins and thus treat certain diseases.

The forefront method for protein structure determination is crystallography. For this, proteins are crystallised and the crystals are exposed to X-rays. Upon interacting with the atoms in proteins, X-rays are diffracted and this can be recorded as an image. From this image we can deduce the atomic structure. However, the process of going from an image to an atomic structure is complex and makes use of many computational analysis tools in order to correctly interpret the data.

As protein structure determination is a chemical problem and we use computational tools to interpret X-ray crystallography experiments, one would expect these studies to belong to a field called “computational chemistry”. However, the term computational chemistry usually denotes computational studies performed after the atomic structure is known, for example to predict affinities of drugs to target proteins or to understand how an enzyme works. Computational chemistry uses physics-based knowledge, such as quantum mechanics, in order to conduct these studies.

Even though computational chemistry and macromolecular crystallography are different fields, they share many methods and problems, Thus, combining computational chemistry and crystallography should lead to improved results in both fields. Images can be more easily interpreted if we use physics-based knowledge of how the protein structure should look like at the end and computational chemistry can make more accurate predictions if it uses as much data as possible from crystallography.

This thesis shows the development and application of combined methods for macromolecular crystallography and computational chemistry. Quantum mechanical calculations are employed to refine crystallographic structures in order to obtain more accurate geometries for interesting parts of the protein. Additionally, information on atomic dynamics present in crystallographic structure is used directly in simulations and the distribution of water molecules that are present around the proteins are improved using computational methods.

List of abbreviations (in alphabetical order)

ADP	Atomic displacement parameter
CASSCF	Complete active space self-consistent field
CRD	Carbohydrate recognition domain
Cryo-EM	Cryogenic electron microscopy
DFT	Density functional theory
DHAP	Dihydroxyacetone phosphate
E&H	Engh & Huber (target values)
Galectin-3C	Carbohydrate recognition domain of galectin-3
GGA	Generalised gradient approximation
GTO	Gaussian type orbital
HF	Hartree–Fock
ITC	Isothermal titration calorimetry
LCAO	Linear combination of atomic orbitals
LDA	Local density approximation
LPMO	Lytic polysaccharide monooxygenase
MC	Monte Carlo (simulations)
MD	Molecular dynamics
MM	Molecular mechanics
MPD	Molybdopterin with deprotonated phosphate
MPH	Molybdopterin with protonated phosphate
MPO	Oxidized molybdopterin
MPT	Molybdopterin
NMR	Nuclear magnetic resonance
PGA	2-phosphoglycolate
PGH	phosphoglycolohydroxamate
PDB	Protein data bank
pMMO	Particulate methane monooxygenase
QM	Quantum mechanics
QM/MM	Quantum mechanics/molecular mechanics
RSCC	Real-space correlation coefficient
RSZD	Real-space Z-difference
SCF	Self-consistent field
TIM	Triosephosphate isomerase
TLS	Translation–Libration–Screw model

1. Introduction

Proteins are arguably the most important molecules in the human body. Therefore, understanding how they function is crucial for medicine, biology and chemistry. In order to reveal the detailed processes taking place inside proteins, protein structures need to be studied at an atomic level. This allows us to study interactions between biomolecules and rationally develop tools that may change these interactions while not affecting the global function of the protein. For example, knowledge of the atomic structure of a target protein is crucial for developing drugs that strongly bind to it and thus help treat diverse diseases.

X-ray crystallography is currently the dominating technique for structure determination of proteins. Due to technological advances, X-ray crystal structures can be rapidly obtained once the proteins are crystallised and the resulting data is of high enough quality to allow detection of individual atoms in the structures. However, subjecting proteins to X-ray beams does not directly give an atomic structure. The diffraction data need to be processed and interpreted. Furthermore, the diffraction data does not contain all information needed to obtain an atomic model, as phase information is not determined experimentally. Thus, the atomic structure is merely a model that fits the data. Protein models contain thousands of atoms, which means the data-to-parameter ratio is rather low in protein crystallography. Because of this, the interpretation of the data does not rely solely on the measurements, but also on some *a priori* chemical knowledge. This chemical knowledge is integrated into the interpretation in the form of restraints, which are very similar to the energy function (the force field) used in computational chemistry to calculate various chemical properties. Modelling the information on protein dynamics, which is present in crystallographic data, may also benefit from using methods employed in computational chemistry to study protein dynamics.

Atomic structures are paramount to computational chemistry. All computational studies in chemistry at the atomic scale need to start from an initial structure and the results depend on the structure chosen. This is especially true in biochemistry, where molecules are too large for an exhaustive sampling of all conformational space. Thus, X-ray crystallography is a very important method for computational chemists as well. However, most scientists in computational chemistry use the deposited protein structures as if they are showing the experimental truth, without

delving into the experimental data and investigating the electron density maps to understand the limitations of the model. This is especially problematic for old structures, for which modern processing techniques had not been used, but also when studying more disordered regions of a protein, which have weak experimental data. Additionally, often the information on atomic dynamics in the protein is ignored when starting a computational study, even if the study is directed towards atomic dynamics.

Unfortunately, there is a gap between macromolecular crystallography and computational chemistry, although they have a lot of methods and problems in common. Crystallography would benefit from using more advanced methods from computational chemistry, whereas computational chemists would benefit from using more than the atomic coordinates when conducting their studies. Most importantly, improved communication between scientists in the two fields may lead to the development of novel methods and the ability to acquire faster and more accurate results. This process has already started, e.g. with combined crystallographic refinement using two of the most widely used software in refinement and molecular dynamics simulations, Phenix and Amber.[1]

The aim of this thesis is to combine methods from protein crystallography and computational chemistry, both in applied studies of enzymatic reactions and from a more methodological perspective. In particular, I have used quantum-chemical calculations in crystallographic refinement, I have attempted to extract more information on atomic dynamics from crystal structure and I improved the modelling of water molecules. This should represent a small step towards bridging the gap between macromolecular crystallography and computational chemistry.

In the following, I first present some basic theory behind macromolecular crystallography and computational chemistry, then delve into more advanced methods that combine both fields, before describing the proteins studied. Finally, I give a brief summary of each of the scientific articles included in this thesis.

2. Crystallography

Macromolecular crystallography is the forefront method for determining atomic structures of proteins, along with nuclear magnetic resonance (NMR) and cryogenic electron microscopy (cryo-EM). Over 90% of the entries in the Protein Data Bank (PDB)[2] are obtained using X-ray crystallography. The first protein crystal ever obtained was of haemoglobin, in 1851, making protein crystallography a rather old method. However, it took another century for the atomic structure of a protein to be determined, Kendrew solving the myoglobin structure in 1958.[3]

In this chapter, I will briefly describe the workflow of macromolecular crystallography and focus especially on the data processing and refinement steps, which share many methods with computational chemistry. The methods described herein will be the ones employed in X-ray crystallography, as this is the most common experimental method in macromolecular crystallography. However, a special section is dedicated to neutron crystallography, as four papers included in this thesis deal with protein neutron structures.

2.1 Overview

All protein crystallography involves five steps: protein crystallisation, data collection, data processing, phasing, and model refinement.

Crystallisation is often the most time-consuming step of obtaining an atomic structure of a protein, as it requires large amount of pure protein and many crystallisation conditions typically need to be screened before a protein crystal is obtained. A more detailed discussion of protein crystallisation is out of the scope of this thesis, and I refer to the book by Ducruix and Giege[4] for the methodology behind protein crystallisation.

Once a good quality protein crystal is obtained, it needs to be exposed to X-rays for experimental data acquisition. The data collection process is usually done at a synchrotron, a powerful source of X-rays, which are produced by high-energy electrons circulating in a storage ring. Improvements over the years have reduced

the data-collection time to only a few seconds, which makes it possible to collect hundreds of data sets in one day.

To collect data, crystals are exposed to X-rays and the diffraction pattern is recorded on a detector. A diffraction pattern is obtained by measuring the intensity of scattered waves as a function of scattering angle. Peaks are obtained in the diffraction pattern when the scattered waves satisfy Bragg's Law:

$$n\lambda = 2d\ sin\theta$$

where θ is the angle of incident and reflected X-rays, λ is the wavelength of the X-rays, d is the distance between the planes and n is an integer representing the order of reflection Thus, each dot in the diffraction pattern forms from the constructive interference of X-rays passing through a crystal. Initially, detectors were X-ray sensitive films, but nowadays pixel-based electronic detectors are used, which are highly sensitive so they can record very weak signals. The diffracted beams are recorded as spots on the detectors and the spacing between the spots indicates the size of the unit cell. The diffraction intensity of each spot represents the actual experimental data, which needs to be processed and interpreted. Data is usually collected from cryo-cooled crystals, at 100 K, although recent advances made possible the collection of high resolution data also at room temperature.

To determine the experimental intensities, the image data needs to be reduced and curated. Data processing in X-ray crystallography includes the following four steps: indexing, integration, scaling and merging.

The lattice symmetry and unit cell size can be deduced from the location of the spots, which in turn makes it possible to do indexing. Indexing represents the process of assigning each spot to a reflection from a specific crystal plane, *i.e.* giving each spot a certain Miller index *h, k* and *l*, in reciprocal space. Each indexed spot is then integrated, to account for background radiation, and scaled. Spots with the same index that appear on multiple images are merged in order to get the total intensity. A diffraction cut-off is applied in order to discard reflections that have a too low signal-to-noise ratio. The value of this cut-off is the highest resolution shell in which reflections can be observed and is called the resolution of the structure. This should result in a data set that is complete up to its highest resolution shell, without radiation damage and correctly scaled and merged. Several statistical terms can be calculated to give information about the quality of the data, such as the correlation coefficient for two halves of the data set, $CC_{1/2}$ or the redundancy-independent residual $R_{meas,}$ which measures the precision of individual intensities.[5,6]

As mentioned above, the intensity of each spot on the detector represents the experimental data. The intensities can also be expressed as structure factors, which describe the amplitude and phase from a set of lattice planes. Intensities are proportional to the square of a structure factor. Each structure factor is a summation of the scattering contribution from each atom in the unit cell:

$$F_{hkl} = \sum_j f_j e^{2\pi i[hx+ky+lz]}$$

where F_{hkl} is the structure factor, f is the atomic scattering factor of the atom *j*, *x, y* and *z* are the Cartesian coordinates in real space, while *h, k* and *l* are the Miller indices.

To relate the experimental structure factors, which are in reciprocal space, to an atomic model, which needs to be built in real space, the electron density can be calculated using a Fourier transform:

$$\rho_{xyz} = \frac{1}{V}\sum_{hkl} |F_{hkl}|\, e^{-2\pi i[hx+ky+lz]} - \phi_{hkl}$$

where ρ_{xyz} is the electron density at position *xyz*, *V* is the volume of the unit cell and ϕ is the phase. Unfortunately, phase information is not contained in the reflected spots on the detected image and thus electron density maps cannot be considered as entirely experimental information. This is what is called the "phase problem" in macromolecular crystallography and can be addressed in a number of ways. The most common method is to use a previously obtained homologous protein structure as a template in order to build the first electron density maps (molecular replacement). This means that the phases are biased by the atomic model. More accurate phases can be obtained if the protein contains heavy atoms such as metals or sulfur atoms from sulfur-containing amino-acids, which present anomalous diffraction due to X-ray absorption. Phases obtained from anomalous diffraction are considered experimental phases, but usually require data collected at multiple X-ray wavelengths (multiple anomalous diffraction). If only a single-wavelength data set is present, single anomalous diffraction phasing can be performed, but this does not provide an unambiguous solution to the phase problem and the ambiguity has to be broken by density modification, for example.[7]

After an electron density map is calculated and a first approximate atomic model built into it, the phasing step is complete. The model (and implicitly, the phases) will then be iteratively refined. The refinement process is described in detail in section 2.3.

2.2 Dynamics in crystal structures

The reflections obtained when collecting data contain information not only about the positions of each atom in the model, but also some information on the dynamics of the atoms in crystal. Although data is usually collected at 100 K, atoms still move at that temperature and since the crystals are flash-cooled, some room-temperature dynamics is frozen in. Dynamic information of e.g. loops around a binding site of is very important for the functional behaviour of the protein. Additionally, room temperature data can also be collected nowadays, which gives even more information on atomic dynamics.

Atomic dynamics in crystal structure can be expressed through two distinct measures: atomic displacement parameters (ADPs), which are also called B-factors or temperature factors, and alternative conformations.

2.2.1 B-factors

The crystallographic B-factors provide a measure of how much atoms vibrate around their positions, thus describing the thermal motion of each atom. They assume a purely harmonic motion, which is not entirely correct, but is considered a good approximation. Furthermore, for resolutions lower than 1.3 Å, only isotropic B-factors are modelled, in which case it is assumed the motion is equal in all directions and can be expressed as:

$$B_j = 8\pi^2 U_j^2$$

where U_j^2 is the mean-square displacement of the atom from its average position. Thus, one can estimate how much an atom vibrates around its position from the B-factor.

For high-resolution crystal structures, anisotropic B-factors can be modelled. These give information on the preferred directions of vibration for each atom and require six parameters per atom. Because of this, using anisotropic B-factor greatly decreases the data-to-parameter ratio, sometimes causing overfitting. An anisotropic B-factor is expressed as a 3 × 3 symmetric tensor:

$$U = \begin{matrix} U^{11} & U^{12} & U^{13} \\ U^{12} & U^{22} & U^{23} \\ U^{13} & U^{23} & U^{33} \end{matrix}$$

The six independent components of the tensor, U^{ij}, are anisotropic displacement parameters that describe how much the atom moves in each direction, thus describing an ellipsoid rather than a sphere for the isotropic B-factor.

B-factors can help to identify regions of the protein that are more mobile or very rigid. However, regions with a high variance in B-factors can also signify errors in the model.

2.2.2 Alternative conformations

Alternative conformations model atoms or groups of atoms that do not occupy the same position in every unit cell, in every asymmetric unit or in every molecule from an asymmetric unit. Physically, this shows that a group of atoms can be found in multiple conformations which represent local minima that can be functionally important, e.g. when a loop opens for the binding and unbinding of a ligand.

Whereas every atom in the molecule will possess a B-factor, not all atoms have alternative conformations. These usually need to be built in manually by the crystallographer after visual inspection of the electron density maps. This means that the information on dynamics contained in a certain atomic structure is subjective and depends on the crystallographer who has built the model. A recently developed software, qFit,[8] can automatically generate alternative conformations starting from a given model and electron-density map. Each alternative conformation of an atom or a group of atoms has an occupancy, which is a number between 0.0 and 1.0 representing the fraction of unit cells a certain conformation is found.

2.2.3 Problems with the dynamics parameters in crystal structures

In theory, B-factors and alternative conformations measure distinct types of dynamics in the crystal and therefore are uncorrelated. However, in practice, B-factor values greatly depend on how the alternative conformations have been modelled. If no alternative conformations have been modelled for a group of atoms that exists in distinct conformations, the static disorder will be partly absorbed by the B-factor, resulting in too high and unreliable values. Additionally, B-factors and occupancy values are coupled and cannot be refined individually.

To account for some of the static disorder of large groups within the protein, one can implement a Translation–Libration–Screw (TLS) model on top of the B-factors. In a simple approximation, TLS models different groups in the protein as rigid bodies that undergo the same movements with the same amplitude.[9] This not only absorbs some of the static disorder included in B-factors, but also helps model anisotropic movements of large groups in cases where individual anisotropic B-factors are not justified (at intermediate to low resolution).

Furthermore, it is sometimes difficult to decide which movement is better modelled by a vibration or by two local minima, as both options are

approximations. An alternative way of modelling dynamics in crystal structures, called ensemble refinement, is presented in detail in section 4.2 and in paper VII.

2.3 Refinement

After the initial phasing and model building, the model is in general not perfect and needs to be improved. Therefore, the last step of determining an atomic structure is model refinement. This is achieved through iterative adjustment of the model parameters until the calculated structure factors from the model, F_{calc}, agree with the observed structure factors from the experiment, F_{obs}, as closely as possible. The model parameters include the atomic coordinates, B-factors and occupancies, but also non-atomic parameters, such as bulk solvent and crystal anisotropy.

In practice, refinement is a minimisation of a target function that is defined such that its value decreases as the model improves. Refinement of each type of model parameters can be done either separately, with a slightly different target function, or all at the same time. Separate target functions used in Phenix are described in the following part. For example, the coordinate refinement target function is expressed as:

$$T_{xyz} = w_x * T_{experiment} + T_{xyz-restraints}$$

where w_x is a weight factor which needs to be determined before refinement, $T_{experiment}$ is the crystallographic term that relates the experimental data to the model structure factors. It can be either a least-squares target or a maximum-likelihood target, the latter being used for most refinements. This target function can be in reciprocal space, i.e. calculated structure factors are compared to experimental structure factors, or in real space, where calculated electron densities are compared against experimental electron densities. As previously described, experimental electron densities are not fully experimental because of the phase problem, so the reciprocal space target function usually yields better results. $T_{xyz-restraints}$ is a restraint term that introduce *a priori* knowledge to the refinement. This term is needed because there are usually insufficient experimental data to account for all the model parameters. This restraint term is defined as deviations from tabulated target values for the geometry of the protein. Target values have been determined by Engh & Huber (E&H)[10] and include bond lengths, angles, dihedral angles and improper dihedral angles, which depend on the chemical nature of the group in question. The restraints term also includes non-bonded, Van der Waals interactions, so that the restraint term can be expressed as:

$$T_{xyz-restraints} = T_{bonds} + T_{angles} + T_{dihedrals} + T_{VdW}$$

The stereochemical E&H restraints have been improved such that the target values for bonds, angles and dihedral do not only depend on the chemical nature of the group, but also on the conformation the group is found in. This new restraint library is called the conformation-dependent library (CDL)[11] and is used nowadays in refinement programs. Stereochemical restraints for unusual molecules (e.g. ligands) need to be calculated before starting the refinement, usually using quantum-chemical methods. Thus, the stereochemical restraints are similar to a molecular-mechanics force field, used in molecular simulations, but they are typically based on a statistical analysis of crystal structures, rather than based on a consideration of energies.

Similarly to the coordinate target function, the B-factor target function is expressed as:

$$T_{bf} = w_{bx} * T_{experiment} + T_{bf-restraints}$$

where $T_{bf-restraints}$ is another restraint term that is meant to keep B-factors of bonded atoms similar. Other restraints can also be included, such as a TLS model. Note that the weight of the experimental target function for B-factor refinement may be different than for the coordinate refinement.

Occupancies are usually only refined for atoms that have an occupancy less than 1.0 (i.e. atoms in alternative conformations) and they are often constrained so that atoms of a group have the same occupancy and so that the occupancies of the different conformations of the same group sum up to unity.

Hydrogen atoms are normally not visible in X-ray crystallography and thus they are not taken into account during the model building and refinement steps. At very high resolutions (<0.8 Å) hydrogen atoms can be modelled individually, otherwise they can be ignored or modelled as riding hydrogen atoms. The latter means that their position depends on the position of the heavy atom they are bonded to and so it is not a refinement parameter. However, their small contribution to the electron density and to F_{calc} is taken into account, which often is a better model than ignoring hydrogen atoms completely.

Protein crystals also contain a large amount of solvent molecules (water), which are mostly disordered. To correctly model the contribution of the solvent to the X-ray scattering in the total model structure factors, a bulk-solvent mask is used:

$$F_{sol} = k_{sol} \exp\left(-\frac{B_{sol} s^2}{4}\right) F_{mask}$$

where k_{sol} and B_{sol} are bulk-solvent model parameters known *a priori,*[12] $s^2 = h'Gh$, where G is the reciprocal space metric tensor, h is a column vector of the Miller indices and h' its transpose. F_{mask} are structure factors calculated from a solvent mask, a binary function that has zero values in regions of the protein and non-zero values in regions with solvent. The mask calculation parameters are also refined in each cycle of refinement.

Water molecules that are well-ordered are visible as oxygen atoms in the density maps. They can therefore be modelled individually and are not included in the bulk-solvent mask. These water molecules can be automatically updated during refinement or built manually after visual inspection.

Refinement can be performed in a variety of software packages, including Phenix,[13] Refmac[14] and Buster.[15] All refinement programs output a coordinate file in the PDB format and a series of electron density maps in order to evaluate the fit of the structure to the experiment visually. The most commonly used maps are the likelihood-weighted $2mF_o - DF_c$ and $mF_o - DF_c$ maps. The former shows electron density around the model, using an extra ($F_o - F_c$) term to minimise model bias, whereas the latter, also called the difference density map, shows regions where the calculated structure factors do not agree with the experimental structure factors.

After visual inspection and model rebuilding in regions with high difference density, a new round of refinement should be run. This process is typically repeated multiple times, as the calculated maps are not independent of the model (because of the phase problem) and need to be recalculated each time the model is changed. This makes refinement a rather time-consuming task and the decision of when a model is complete quite subjective.

2.4 Model validation

To prove that the refined model is correct, some validation metrics are needed. Two types of metrics can be calculated: global quality metrics, which are statistical measures that show if the whole model agrees with the experimental data or with chemical knowledge, and local quality metrics, which shows how well atoms or groups of atoms fit in the electron density.

2.4.1 Global quality metrics

The most common measure to track agreement of the model to the data is the *R* value. The *R* value is expressed as:

$$R = \frac{\sum \left| |F_{obs}| - |F_{calc}| \right|}{\sum |F_{obs}|}$$

An ideal *R* value is 0, but the value depends on the resolution of the data. A desirable target *R* value for a 2.5 Å resolution data set is around ~0.2. Two different R values are provided by refinement programs, one that is based on all the reflections in the data, also called R_{work}, and one that is based on a small set of reflections (5 to 10%), chosen randomly and not used in the refinement, which is called R_{free}. If the difference between these two *R* values is very large the model is considered to be overfitted and additional refinement needs to be performed.

Apart from agreement to the data, a model should also agree to prior chemical knowledge, e.g. giving a reasonable geometry of the atoms. All PDB entries report the root mean square deviation of the bond lengths, angles and dihedrals from ideal tabulated values and the percentage of protein torsion angles outside allowed areas in a Ramachandran plot.[16]

2.4.2 Local quality metrics

To be able to fully interpret regions of the protein model that are interesting from a biological point of view, we need to make sure that the specific region is correctly modelled, which cannot be deduced from global quality metrics alone.

The most common local quality metric is the real space correlation coefficient (RSCC), which is a measure of the similarity between an electron density map calculated directly from a structural model and one calculated from experimental data. It is expressed as:

$$RSCC = \frac{\sum |\rho_{obs} - < \rho_{obs} >| \sum |\rho_{calc} - < \rho_{calc} > |}{\sqrt{(\sum |\rho_{obs} - < \rho_{obs} >|^2 \sum |\rho_{calc} - < \rho_{calc} > |^2)}}$$

The RSCC can be calculated for any arbitrary set of atoms, such as one protein residue or one functional group. An ideal RSCC value is 1.0. As the R values, the RSCC reports on both accuracy and precision.

Another local quality metric is the real-space Z-difference (RSZD). It is normally considered the best measure to locally evaluate the goodness-of-fit for a group in a crystal structure and essentially evaluates the largest and smallest values in the $F_o - F_c$ difference density map around the group of interest. Additionally, the RSZD is a measure of just the accuracy of the model and not the precision. RSZD can take both negative and positive values, with the negative values suggesting that there is too much electron density in the model in region studied (e.g. an atom has been modelled where there should not have been any), whereas a positive value indicates that there is too little electron density in the model in the region

studied. An ideal RSZD value is 0.0, but in general, regions with an absolute RSZD lower than 3.0 are considered to have a proper fit to the data.

Finally, B-factors can be used as local quality metrics, at least in a qualitative manner. If a group of atoms has B-factors that are much higher than the average B-factor of the protein, it is likely that it has been incorrectly modelled.

2.5 Neutron crystallography

Hydrogen atoms play an important role in the function of many protein, especially enzymes, as there exist many enzymatic reactions which involve proton (H^+) transfers. However, as previously mentioned, X-ray crystallography is not able to detect the atomic positions of hydrogen atoms, because they contain only a single electron that interacts with the X-rays. In contrast, neutrons interact with the nuclei and the diffraction is different for each isotope. Unfortunately, hydrogen (^1H) atoms have a negative scattering length and also a high incoherent scattering cross-section in neutron experiments. However, deuterium (^2H) atoms give a coherent scattering, similar to that of carbon or oxygen atoms and therefore, their positions are visible in neutron maps.

Thus, if a sufficient number of hydrogen atoms are exchanged with deuterium in the protein structure, neutron experiments can be performed in order to elucidate the positions of hydrogen atoms inside the protein. Furthermore, neutron diffraction experiments cause no radiation damage, so the crystal used for acquiring the neutron data can be afterwards employed in the complementary X-ray diffraction experiments. Unfortunately, the brilliance of neutron sources is rather low, so large crystals are needed and experiments are costly and takes several days or weeks to complete.

Data processing for neutron crystallography follows the same steps as for X-ray crystallography: indexing, integration, scaling and merging. Refinement of protein neutron crystal structures also behaves exactly the same as X-ray refinement, with the exception that a different scattering library has to be used in order to calculate the structure factors and hydrogen (and deuterium) atoms must be treated individually.

However, neutron data are often of a lower resolution than X-ray data from the same protein. To make matters worse, the model parameters that need to be refined are almost doubled due to the inclusion of hydrogen atoms in the refinement, which can often cause overfitting. Fortunately, as X-ray experiments can be conducted on the same crystal as neutron experiments, X-ray data can be included in the refinement along with the neutron data, thus increasing the data-

to-parameter ratio. This refinement procedure is called joint X-ray/neutron refinement and follows the same procedure as traditional refinement, but the target functions contain both neutron and X-ray experimental functions. For example, the coordinate target function is expressed as:

$$T_{xyz} = w_x * T_{X-ray} + w_n * T_{neutron} + T_{xyz-restraints}$$

Different weights are used for the X-ray (w_x) and neutron (w_n) experimental functions and thus the model can be biased to either data set, depending on their quality.

We use protein neutron crystal structures and joint X-ray/neutron refinement in papers IV–VI and X of this thesis.

3. Computational chemistry

In this chapter, I will present the basic theory behind several approaches of studying chemical systems computationally. Quantum mechanical (QM) methods are expensive but give accurate results, whereas molecular mechanical (MM) methods are cheap, but do not treat electrons explicitly, resulting in less accurate results, especially for systems where the electronic structure is important (e.g. systems containing metal atoms). MM methods are especially suitable for sampling many conformations of the studied system, which is usually done through a molecular dynamics (MD) simulation. A good trade-off for studying biomolecular systems is using the combined QM/MM method, which I also present herein.

3.1 Quantum mechanics

QM methods have their origins in the beginnings of the 20th century and are largely based on the time-independent Schrödinger equation:

$$\hat{H}\Psi = E\Psi$$

where $\hat{H}$ is the Hamiltonian operator, Ψ is the wavefunction and E is the total energy of the system. The Hamiltonian operator consists of a kinetic energy term and a potential energy term, each of these being separated into nuclear and electronic terms, respectively.[17] According to the Born–Oppenheimer approximation, the nuclear and electronic terms can be treated separately, because electrons are much faster and lighter than nuclei and thus the atomic nuclei can be considered stationary with respect to the electrons.

3.1.1 Hartree–Fock approximation

The Schrödinger equation cannot be exactly solved for systems more complex than the hydrogen atom. Therefore, multiple ways to approximately solve the Schrödinger equation have been developed. The simplest of these is the Hartree–Fock (HF) approximation,[18] which is a mean-field approximation, i.e. each electron is considered in the average field generated by the other electrons. In this

way, the total *N*-electron wavefunction can be written as a Slater determinant, an antisymmetrised product of one-electron wave functions.

Another approximation used in HF is the linear combination of atomic orbitals (LCAO). This approximation expresses molecular orbitals ψ as a linear combination of all atomic orbitals χ_i with different coefficients c_i:

$$\psi = \sum_i c_i \chi_i$$

The HF equations are solved iteratively, starting from an initial guess, which is refined until the solution reaches convergence. Due to this solving algorithm, HF is also known as the self-consistent field (SCF) formalism.

The HF method neglects the correlation of electron motion, which can lead to large errors in the calculation, so HF shows limited usage in modern quantum chemistry calculations. Many post-HF methods have been developed to include electron correlation, either in a dynamic form through perturbation theory, e.g. MP2 or in a static form (through multiple electronic configurations), such as complete active space SCF (CASSCF). The detailed description of post-HF methods is out of the scope of this thesis.

3.1.2 Basis sets

Basis sets are mathematical functions that are used to construct molecular orbitals (for example, through the LCAO approximation). The most commonly used basis sets in computational chemistry are Gaussian-type orbitals (GTOs),[19] i.e functions of the type:

$$\chi_{ijk,\propto} = x^i y^j z^k e^{-\propto r^2}$$

Basis sets are usually constructed from a linear combination of contracted GTOs. The number of contracted GTOs used for each electron pair gives the dimension of the basis set: split-valence (one basis function for inner shell electrons and two for valence electrons), double-zeta (two basis functions for each electron pair), triple-zeta (three basis functions for each electron pair), etc.

Furthermore, other functions can be added to the basis functions to improve accuracy. GTOs of greater angular momentum can be added as polarisation functions to certain atoms. These allow orbitals to be asymmetrical, which is important to describe the polarisation of chemical bonding. Diffuse functions are GTOs with small exponents, which are needed when describing negatively charged species.

In principle, an infinite basis set would give a perfect accuracy, but that can of course not be obtained in practice. As described above, the optimal basis set for quantum mechanical calculations is system dependent and must be carefully chosen for optimal balance between accuracy and computation time. In practice, accurate molecular geometries can be obtained with a rather small basis set, whereas accurate energies require a bigger basis set.

3.1.3 Density functional theory

Density functional theory (DFT) is the most common QM method used at present in computational chemistry, due to its speed and accuracy.

DFT has its roots in solid-state physics, where it has been used since the 1970s. The basic concept of DFT relies on considering the electron density, which is a function of only the three Cartesian coordinates: x, y and z, rather than the wavefunction in the Schrödinger equation, which is a function of $3N$ variables.[20] The Hohenberg–Kohn theorem proves that all ground-state properties of a system can be derived from the electron density, thus making DFT a useful tool for quantum chemical calculations. The energy functional (function of a function) used in DFT can be divided into the following components:[21]

- The kinetic energy of non-interacting electrons
- The classical interelectronic repulsion
- The nuclei–electrons attraction
- The correction to the electrons' kinetic energy due to interactions
- Non-classical corrections to the interelectronic repulsion energy

The first three components can be computed exactly, whereas the last two are unknown and usually combined together in a term called exchange–correlation energy. However, the exchange–correlation energy includes dynamic electron correlation, which is one of the shortcomings of HF. There are several ways to approximate the exchange–correlation energy. The local density approximation (LDA)[22] assumes that the density can be treated locally as a uniform electron gas. The generalised gradient approximation (GGA) expands the LDA by including the first derivative of the density in the energy expression and is successfully used in many widely-used functionals (e.g. PBE,[23] BLYP[24]). Meta-GGA functionals (e.g. TPSS[25]) take this one step further and also include the second derivative of the density. Owing to the improper treatment of exchange, some functionals, called hybrid functionals, include a certain amount of exact HF exchange[26] in their energy term. B3LYP,[27] arguably the most popular functional today, is a hybrid functional.

Although it gives accurate and fast results compared to HF, DFT has its own share of problems. Firstly, it is not trivial to choose which functional to use when starting a calculation. Usually, multiple functionals are tested, and a more complex functional is not guaranteed to give better results. Moreover, DFT (like HF) cannot treat non-bonded dispersion interaction. However, the parametrised DFT-D methodologies proposed by Grimme[28] provide a fast and accurate correction to DFT for dispersion interactions.

3.2 Molecular mechanics

In contrast to QM methods, MM methods do not explicitly take into consideration the electronic structure of the molecules. Instead, MM considers only the positions of the atomic nuclei and the bonds between the atoms, in a so-called "ball and spring" model of a molecule. The energies of molecules are calculated with the aid of force-fields.

Force-fields are potential energy functions that relate the energy of a molecule to the positions of all atoms in it. A typical force-field used in biomolecular computational chemistry consists of a number of physical terms:

$$E_{total} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{VdW} + E_{el}$$

The first three terms describe the internal energy of the molecule as a function of the bonds, angles and dihedrals present in it. The last two terms are non-bonded terms, describing Van der Waals interactions and electrostatic interactions, respectively. One can notice that the terms in the force-field energy function are very similar to the stereochemical E&H restraints used in crystallographic refinement, except for the electrostatic term, which is usually ignored in crystallographic refinement. Therefore, technically any force-field can be used as geometry restraints during refinement. For example, Amber force-fields have been implemented in the Phenix software.[1]

The individual terms are mathematically expressed as follows:

$$E_{bonds} = \sum k_b(r - r_0)^2$$

where k_b is a spring force constant and r_0 is the ideal bond length for the given type of bond. This is a harmonic potential and it is also used to describe the bond angle term:

$$E_{angles} = \sum k_a(\theta - \theta_0)^2$$

The dihedral term uses a periodic function:

$$E_{dihedrals} = \sum \frac{V_n}{2}(1 + \cos(n\phi - \delta))$$

where V_n is a force constant, n is the periodicity of the torsion angle ϕ, while δ is the phase shift.

The Van der Waals interactions are normally described by a 6–12 Lennard-Jones potential, which models both attractive (dispersion) interactions and repulsive (exchange) interactions, such that the energy is repulsive at very short interatomic distances, but slightly attractive at intermediate distances:

$$E_{VdW} = \sum_i \sum_{j \neq i} 4\epsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}^{12}} - \frac{\sigma_{ij}}{r_{ij}^{6}}\right)$$

where ϵ_{ij} is the depth of the potential energy curve and σ_{ij} is the distance where $E_{vdW} = 0$.

The electrostatic term can be described as a classical Coulomb term:

$$E_{el} = \sum_i \sum_{j \neq i} \frac{q_i q_j}{4\pi\epsilon\epsilon_0 r_{ij}}$$

where ϵ_0 is the vacuum permittivity, ϵ is the relative permittivity of the medium, while q_i is the partial charge of the atoms.

As can be seen from the equations, MM methods are heavily parametrised. All force constants, ideal bond, angle and dihedral angles, as well as all Van der Waals parameters and atomic charges need to be obtained before a calculation can be performed. The parameters in the various existing force-fields are either obtained from experimental data or from high-level QM calculations. Many force-fields have been developed for protein studies, with each being optimal for different kinds of systems (e.g. AMBER,[29] CHARMM,[30] Gromos[31] and OPLS[32]). Parameters for unusual residues (e.g. ligands, co-factors) that are not included in the force-field library, need to be derived separately.

MM cannot be used to study chemical reactions as the bonds are defined by the user prior to the calculation, so that no bond-breaking can occur. However, MM methods can be used for geometry minimisation or energy calculations, like the QM methods, but are most useful for sampling many conformations of a specific

system because of their low computational cost. In computational biochemistry, sampling is usually done through molecular dynamics simulations, which are discussed in section 3.4.

3.3 QM/MM

To take advantage of the accuracy of QM methods and the low computational cost of MM methods, a combined approach has been developed, called QM/MM, introduced by Warshel and Levitt.[33] In this method, a small part of the system (also called quantum system or system 1) is treated with QM methods, whereas the rest of the system is treated by MM methods. The method has also been generalised so you can use more than two levels of theory (e.g. DFT, semi-empirical and MM), through the ONIOM formalism.[34] QM/MM methods are especially useful when studying enzymatic reactions, as the molecular part where the reaction takes place can be included in the quantum system, which allows bond-breaking to occur.

There are several ways to couple the QM and MM regions. In the subtractive scheme, the total energy of the system is calculated at the MM level and the QM energy of the quantum system is calculated. To avoid double-counting, the MM energy of the quantum system is then subtracted from the sum:

$$E_{QM/MM} = E_{MM12} + E_{QM1} - E_{MM1}$$

An additive scheme can also be used, where MM energies are only calculated for system 2, together with an interaction energy between the two systems:

$$E_{QM/MM} = E_{MM2} + E_{QM1} + E_{QM/MM}$$

The electrostatic interactions between the QM and the MM systems can also be described in different ways. In mechanical embedding, all interactions are treated at the MM level. This can be problematic, as MM methods do not take into account electronic polarisation.

To improve this, one can use an electrostatic embedding scheme, in which MM point charges are included in the QM calculation, thus polarising the quantum system. In the subtractive scheme of electrostatic embedding, the charges of the quantum system are zeroed at the MM level in order to avoid double-counting.

In most cases, there are covalent bonds between the quantum system and the MM system, which need special treatment. Several ways of treating covalent bonds between the systems have been developed.

The easiest solution is to introduce an extra truncating (link) atom in the quantum system, which is usually modelled as a H atom. This atom is placed and fixed at the position of the heavy atom in the MM system.

Another solution is to replace the chemical bond between the MM and QM systems with a special orbital. This assumes that the nature of the bond is not sensitive to changes in the QM region. Approaches that use this method are the localised hybrid orbital approach[35] or the generalised hybrid orbital approach.[36]

In this thesis, all QM/MM calculations have been performed using the ComQum interface between the Turbomole and Amber softwares.[37,38] This method uses a subtractive scheme with electrostatic embedding and hydrogen link atoms.

QM/MM has been used in papers III, V and VI to study the reaction mechanisms of various enzymes.

3.4 Molecular dynamics

The energy of a molecule depends on the position of its atoms in space. The computational methods discussed so far can calculate the energy of a specific geometry of a molecule or find the molecular geometry with the lowest energy, but cannot generate multiple conformations of the same molecule. For that, a sampling method is needed and a simulation of the molecule must be performed. The most common sampling methods used in computational biochemistry are molecular dynamics (MD) and Metropolis Monte Carlo (MC) simulations. Throughout this thesis only MD simulations have been used to sample the biomolecular systems.

MD simulations employ Newton's second law of motion to move the atoms in the system:

$$F = ma = m\frac{d^2r}{dt^2}$$

where F is the force, m is the mass and a is the acceleration, which can also be written as the second derivative of the position (r) with respect to time (t). From this differential equation, positions of all atoms can be determined by integrating the motion in small timesteps if the force acting on each atom is known. The force is simply calculated as the first derivative of the potential energy with respect to the position of the atom:

$$F = -\frac{dU}{dr}$$

The potential energy function can be calculated either by QM or (more commonly) by a MM force field.

Therefore, an MD simulation iteratively calculates the forces acting on the atoms at a given position, then moves the atoms according to these forces over a certain timestep. A typical timestep used in MD simulations of biomolecules is 0.5 fs, as this is smaller than the period related to fastest vibrational frequency in the system (that of hydrogen atoms). If the bond lengths involving hydrogen atoms are kept fixed (e.g. by using the SHAKE algorithm[39]), slightly larger timesteps (2 fs) can be used.

The length of the MD trajectory (the number of timesteps employed) has to be decided beforehand and depends on the problem one is studying. Typically hundreds of nanoseconds need to be run, with the longest ever protein MD simulation being 1 ms.[40] As the final trajectory depends on the initial positions and velocities of the atoms, a common method to enhance sampling is to start multiple rather short simulations. In this work, all MD simulations were run as 10 simulations of 10 ns.

MD simulations can be made massively parallel and most software packages have implemented the MD algorithms on GPUs. However, obtaining a 1 μs trajectory still requires significant computational time. A widely used method to reduce the computational effort is to employ periodic boundary conditions and calculating electrostatic interactions through particle mesh Ewald summation,[41] which reduces the scale of the simulations from $O(n^2)$ to $O(n \log n)$, where n is the number of atoms.

3.4.1 Preparing the system

Before performing an MD simulation, the system studied needs to be solvated and equilibrated. The first step in the preparation is assigning protonation states to all residues in the protein by studying the hydrogen bonding network or from information from neutron structures. Then, a box of water molecules needs to be added around the protein, which must be large enough to avoid any boundary effects. The box shape can vary (octahedral, cubic or spherical) and there are many types of water models to choose from (TIP3P,[42] SPC/E,[43] TIP4P,[42] etc.). The choice of the size of the water box and of the water model also depends on the scientific problem studied. Then, the atoms of the macromolecule and the solvent need to undergo a relaxation so the system reaches an equilibrium state. This typically takes some nanoseconds.

Instead of solvating the protein, which mimics a protein in solution (e.g. in the human body), one can construct a simulation from the unit cell in which the protein is found in the crystal. This simulates dynamics of the protein within the crystal and is especially useful for comparison with experimental diffraction data.

Although placing the protein in the unit cell is straightforward, solvating the unit cell can cause some issues in system preparation. As the exact number of water molecules in the unit cell is not known, many water compositions need to be tested in short MD simulations and the one that keeps the system volume most stable is chosen for further simulations. Additionally, equilibration of crystal MD simulations usually takes longer than for traditional MD simulations, as the system is less isotropic.

3.4.2 Calculating conformational entropy

Calculating conformational entropy is an important task in computational chemistry, as it is an essential part of the binding free energy of a ligand, along with the other elements that contribute to the total entropy, i.e. solvation entropy and rotational–translational entropy.

Estimation of conformational entropy relies on computational methods, as no experimental techniques can fully measure it. Information on the dynamics of each atom is needed to calculate conformational entropy, so an MD trajectory needs to be computed before entropy calculation.

There are several methods to calculate conformational entropy once one has an MD trajectory. Dihedral histogramming implies investigating the distribution of dihedral torsion angles of the protein residues. In order to do this, Cartesian coordinates are transformed to internal coordinates, then a discrete histogram of the dihedral angles in increments of 1–10º is constructed and the entropy is calculated from:[44]

$$S = \frac{R}{2} - RlnN - R\sum_{i=1}^{N} p_i lnp_i$$

where R is the gas constant, N is the number of bins employed and p_i is the probability that the dihedral angle is found in bin i.

Another common method to estimate conformational entropies is quasi-harmonic analysis. This is implemented in most MD analysis packages, such as *cpptraj* from AmberTools. The atomic fluctuations are assumed to follow a multivariate Gaussian distribution and quasi-harmonic frequencies are calculated as the eigenvalues of the mass-weighted variance–covariance matrix of the atomic fluctuation determined from an MD simulation. The conformational entropy is then calculated from:

$$S = \frac{h\omega}{T}\frac{e^{-\beta h\omega}}{1 - e^{-\beta h\omega}} - k_B \ln(1 - e^{-\beta h\omega})$$

where h is Planck's constant, T is the temperature, k_B is Boltzmann's constant, β is $1/k_B$, and ω is a quasi-harmonic frequency. This equation can also be used by substituting the frequencies obtained from the covariance matrix with vibrational frequencies calculated from normal-mode analysis.[45]

In paper VIII, we investigate if we can use crystallographic information on atomic dynamics for conformational entropy calculation without performing MD simulations.

4. Advanced methods

4.1 Quantum refinement

Quantum refinement is a method that combines crystallographic refinement (see section 2.3) and QM/MM (see section 3.3). As described above, crystallographic refinement requires a set of geometry restraints, usually E&H restraints, in order to obtain reliable bond lengths and angles, and not to overfit the crystallographic data. These restraints are similar to a force field used in a MM treatment of biomolecules. Thus, the geometry restraints can be replaced for a small part of the system (called system 1 as in the regular QM/MM method) by a QM energy function and then added together with the crystallographic data to generate a quantum refinement target function:

$$E_{\mathrm{cqx}} = w_{\mathrm{X}}\, E_{\mathrm{Xray}} + E_{\mathrm{MM12}} + w_{\mathrm{QM}}\, E_{\mathrm{QM1}} - E_{\mathrm{MM1}}$$

where E_{Xray} is the crystallographic data, w_X is the normal X-ray/restraints weight, E_{MM12} is the E&H derived energy for the whole system, whereas E_{QM1} and E_{MM1} are the QM and E&H energies for system 1, respectively. An extra weight factor, w_{QM}, is needed to scale the QM energies, as the E&H restraints are not energy-based and cannot be added or subtracted directly from a QM energy. This weight factor has been found to be 3 in previous studies. The practical details of the interaction between system 1 and the rest of the molecule is the same as in the QM/MM formalism and is discussed in section 3.3.

This method has been implemented in the ComQum-X software,[46,47] which combines the Turbomole QM software[48] with CNS software[49] for crystallographic refinement. Only the coordinate refinement of system 1 is done in Turbomole, using the *relax* program, which employs a Broyden–Fletcher–Goldfarb–Shanno quasi-Newton approach. The other steps of crystallographic refinement are performed with minimal changes in CNS, in order to write out crystallographic energies and forces that are passed to Turbomole. A scheme of the flow of the ComQum-X software is presented in Figure 4.1.

Other schemes of including quantum chemistry in crystallographic refinements have been developed in several software programs. Merz and Westerhoff have implemented linear-scaling semiempirical calculations of the whole protein in the

DivCon program.[50,51] The *Q|R* project also refines whole proteins using quantum chemical methods.[52]

```
Evaluate the QM wavefunction
Repeat
      Evaluate the QM forces (within S1)
      Evaluate the crystallographic forces (from S1 & S2 onto S1)
      Add the forces
      Relax the geometry of S1 using these forces
      Change the coordinates of S1 in CNS representation
      Relax S2 by crystallographic refinement with S1 fixed
      Perform an individual B factor refinement of S1 & S2
      Evaluate the QM wavefunction and energy of S1
      Evaluate the crystallographic energy function
      Add the energies until convergence
Perform an individual B factor refinement of S1
```

Figure 4.1 Flowchart of the ComQum-X program. S1 and S2 denote systems 1 and 2. Steps in bold constitute the ComQum-X interface. Steps in italics are performed by CNS software, whereas underlined steps are performed by the Turbomole software.

Thereby, quantum refinement introduces an accurate energy function for the system of interest. This is especially useful for protein active sites that contain unusual ligands, for which there are no tabulated geometry restraints and the restraints would have to be manually calculated before refinement. Furthermore, the treatment of electrons is essential for finding the correct coordination geometry of metals. Thus, quantum refinement can greatly improve the geometry of the active site of metalloenzymes.

Quantum refinement can be easily extended to joint X-ray/neutron refinement by replacing the geometry restraints in the joint target function with QM/MM restraints:

$$E_{\mathrm{cqu}} = w_{\mathrm{X}}\, E_{\mathrm{Xray}} + w_{\mathrm{N}}\, E_{\mathrm{neutron}} + E_{\mathrm{MM12}} + w_{\mathrm{QM}}\, E_{\mathrm{QM1}} - E_{\mathrm{MM1}}$$

Only a $w_{\mathrm{N}}\, E_{\mathrm{neutron}}$ term needs to be added to account for the neutron data and the neutron data weight, while the rest of the equations remain the same as for regular quantum refinement. Neutron quantum refinement is presented in detail in paper IV.

Quantum refinement is applied to various biomolecular systems in papers I–VI.

4.2 Ensemble refinement

Currently, most of the protein crystal structures deposited in the PDB are static snapshots, with limited information about the underlying atomic dynamics that exist in crystal structures, in the form of B-factors and alternate conformations (see section 2.2). To visualise all the information regarding dynamics that affect X-ray diffraction a different refinement method is needed.

Burnley et al. recently devised a method that combines MD simulations with crystallographic data, called ensemble refinement.[53] This produces an ensemble model in which atomic fluctuations are represented by multiple structures within the ensemble. This is implemented in the *Phenix* software within the *phenix.ensemble_refinement* module.[13]

Earlier attempts at performing MD simulations with crystallographic restraints[54] failed due to overfitting of the models with respect to the data. To overcome this problem and increase the data to parameter ratio, ensemble refinement models large-scale motions by a TLS, global disorder model. This allows the MD simulation to sample only relevant, local atomic fluctuations.

In practice, the MD simulation is performed using time-averaged restraints, resulting in time-averaged structure factors, according to the following equation:

$$< F_{calc} > = e^{-\Delta t/\tau} < F_{calc} >_{t-\Delta t} + (1 - e^{-\frac{\Delta t}{\tau}}) F_{calc}^{t}$$

where Δt is the time-step in the ensemble refinement simulation, typically 4 fs. The size of the averaging window is controlled by the τ parameter that needs to be set before each refinement, but is not system-dependent and is usually set to 1 ps.

The non-solvent atoms are coupled to a Berendsen temperature-bath[55] during simulation, with a temperature that is 2-10 K less than the target temperature (usually 298 K), because the X-ray restraints energy term is non-conservative and causes heating. Every 10 time-steps, the X-ray weight is modulated by the temperature of the protein atoms ($T_{protein}$), such that all atoms sample consistently at the target temperature (T_{target}):

$$w_x^t = \frac{w_x^{t-\Delta t}\, T_{target}}{T_{protein}}$$

In this way, the thermostat offset controls the X-ray weight in order to maintain the sampling temperature constant. The X-ray–temperature bath coupling constant is set before starting the ensemble refinement and controls the strength

of the X-ray restraints. This parameter is also system independent and usually set to 10 K.

The global disorder is modelled with the TLS approximation. This is calculated before the start of the simulation using the B-factors of the traditionally refined structure. The number of TLS groups per molecule must be decided manually as appropriate for the studied system. For each group, TLS parameters are fitted to the starting B-factors iteratively. In each step, a percentile of the atoms with the poorest fitting of the TLS parameters are excluded from the following round and this is repeated until the TLS parameters converge. The selection of the TLS groups and the percentile of atoms to be excluded is non-trivial and system-dependent. Multiple ensemble refinements with different TLS parameters need to be performed before running a final ensemble refinement simulation.

The modelling of the dynamics of water molecules with crystallographic data is a challenge due to the noisy nature of the electron density of water molecules in protein crystal. In ensemble refinement, water molecules are deleted and repositioned in electron density peaks every 250 time steps to avoid protein clashes or unphysical movements of the water molecules.

Ensemble refinement is still a rather new method and only few studies have been conducted that use it as a tool to e.g. understand loop dynamics in an enzyme.[56] In paper VIII, we use ensemble refinement as a complementary technique to quantitative methods such as MD to explore ligand dynamics in the binding site of galectin-3.

5. Systems studied

The methods described previously in this thesis are useful only if they help us to understand the structure and function of proteins in nature. Herein, I describe the proteins I have studied in this thesis using combined computational–crystallographic methods.

5.1 Galectin-3

Galectin-3 is a member of the galectin family of mammalian lectins.[57] For the studies in this thesis, the carbohydrate recognition domain (CRD) of galectin-3 was used, which is denoted throughout the papers, as galectin-3C.[58] The CRD is a conserved sequence motif in the galectin family that confers affinity for β-galactoside containing glycans. These glycans bind in a relatively solvent-accessible binding site, situated in a shallow groove across one of the two β-sheets of galectin-3C (Figure 5.1). Galectin-3 has been found to play a role in various biological processes, such as cell growth, cell differentiation, cell cycle regulation, signalling, and apoptosis, which makes it an interesting pharmaceutical target to fight inflammation and cancer.[59]

Apart from its biomedical application, galectin-3C has been studied for two other reasons: a variety of synthetic ligands with different affinities that bind to galectin-3 has been synthesised and galectin-3C easily crystallises in large crystals, making it a perfect model system for crystallographic studies.

Figure 5.1 Example of Galectin-3C in complex with a synthetic ligand.

Galectin-3C has been used as a model system in papers IV, VII, VIII, IX and X to test for novel methods such as neutron quantum refinement or ensemble refinement.

5.2 Particulate methane monooxygenase

Particulate methane monooxygenase (pMMO) is one of the only two group of enzymes that can oxidise methane, along with the soluble methane monooxygenases.[60] pMMOs are the predominant enzymes in methanotrophic bacteria and are membrane-bound, which makes characterization of these enzymes difficult.[61] Thus, the resolution of the crystal structures of pMMO is rather poor.

Moreover, the location and nature of the active site has been disputed. Different putative active site have been proposed, composed of a dinuclear Fe site or of mono-, di- tri-nuclear Cu site.[62–64] Crystallographic and EXAFS studies have suggested that the active site of pMMO contains three histidine residues and two Cu ions.[62] The histidine residues interact with the metals in a so-called histidine brace, in which one residue is terminal and binds the metal both with the side-

chain N atom and with the terminal N atom. However, the structure is strange and does not make chemical sense (cf. Figure 5.2).

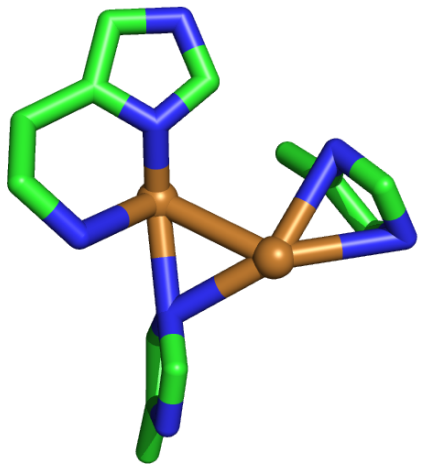


Figure 5.2 Proposed dinuclear copper active site of pMMO in the 3RGB crystal structure.

In paper I, we apply quantum refinement to elucidate the nature and number of metal centres in the active site of pMMO, which cannot be clearly discerned from crystallographic data alone.

5.3 Nitrogenase

Nitrogenase is the only enzyme in nature that can break the triple bond in N_2 molecules, creating ammonia and making nitrogen available to organisms. Understanding the exact catalytic mechanism of nitrogenase could enable us to use a more energy-efficient way of producing ammonia, a chemical that is crucial for agriculture.[65] The full nitrogenase reaction is:

$$N_2 + 8\ e^- + 8\ H^+ + 16\ ATP \rightarrow 2\ NH_3 + H_2 + 16\ ADP + 16\ P$$

The nitrogenase reaction takes place at a complicated FeMo cluster with the composition $MoFe_7S_9C$(homocitrate), which binds to the protein through a histidine and a cysteine residue.[66] (Figure 5.3) The Mo ion is replaced by V and Fe in some variants of the enzyme.[67] The protein also contains two more FeS clusters, which facilitate electron transfer.

A range of diverging reaction mechanisms have been proposed so far for the nitrogenase reaction from experimental and computational studies and no consensus has been reached, except that 4 electrons and 4 protons need to be added to the FeMo cluster before N_2 can bind.[68–70] To elucidate the full mechanism, complex QM/MM calculations need to be performed, which in turn require the protonation state of all atoms in the protein.

Although high-resolution crystal structures have been obtained, information about the protonation states of ionisable residues cannot be obtained from X-ray diffraction alone. The protonation of the homocitrate ligand is especially important, as it coordinates directly to the Mo atom.

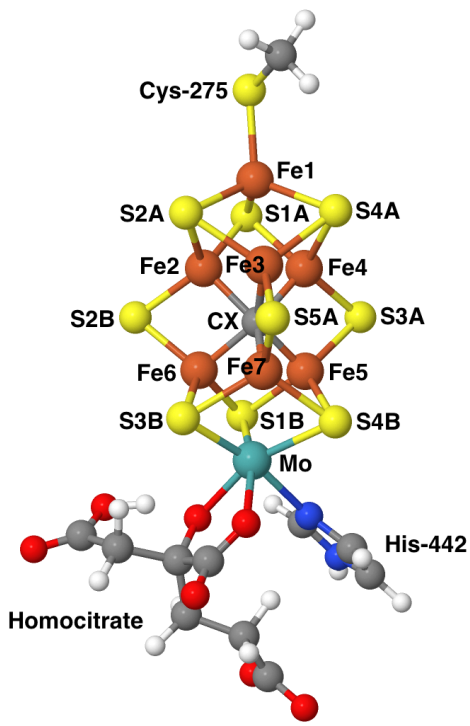


Figure 5.3 The FeMo cluster in nitrogenase according to the 3U7Q crystal structure.

In paper II, we use quantum refinement combined with MD simulations and QM calculations to elucidate the protonation state of the homocitrate ligand in nitrogenase.

5.4 Sulfite oxidase

Sulfite oxidase is another molybdenum-containing enzyme.[71] It catalyses the oxidation of sulfite in many organisms, including humans, in the final step of the degradation of the sulfur-containing aminoacids.[72] The Mo atom in the active site is coordinated to a special ligand, molybdopterin (MPT). It also coordinates a cysteine residue and one or two oxo groups, depending on the oxidation state. In the oxidised structure, the geometry of the active site is square pyramidal, with one of the oxo groups in the axial position.[73] (Figure 5.4)

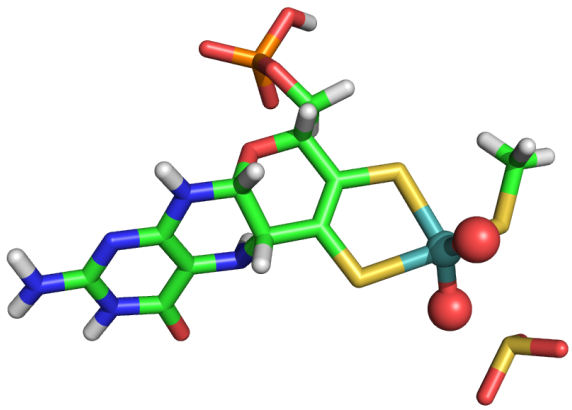


Figure 5.4 Sulfite oxidase active site.

Although the geometry and composition of the active site are well-established, several details about the chemistry of sulfite oxidase are still disputed. Firstly, several reaction mechanisms have been proposed:

- the lone pair of the sulfur atom of sulfite attacks the equatorial oxo ligand of Mo, forming sulfate (S–OMo mechanism)[74]

- sulfite coordinates to the Mo atom with one of its O atoms before sulfate is formed by the attack of the equatorial oxo ligand (O–Mo mechanism)[75]
- sulfite coordinates to the Mo atom with the S atom before sulfate is formed by the attack of the equatorial oxo ligand (S–Mo mechanism).[76]

Secondly, the MPT ligand can exist in several states: with the phosphate group deprotonated (MPD) or singly protonated (MPH) and oxidised in a 10,10a-dihydro state (MPO).[77]

In paper III, we use QM/MM calculations and quantum refinement to decide the mechanism of sulfite oxidase and the state of the MPT ligand.

5.5 Lytic polysaccharide monooxygenase

The lytic polysaccharide monooxygenases (LPMOs) are a family of metalloenzymes that oxidise the C–H bonds of the glycoside link in polysaccharides, by activating molecular oxygen.[78,79] This leads to enhanced polysaccharide decomposition, which in turn could lead to more energy-efficient production of biofuels from common polysaccharides, such as cellulose.[80,81]

The active site of LPMOs is rather similar to one of the proposed active sites of another oxygen-activating enzyme, described earlier, pMMO. It contains a Cu ion coordinated by two histidine residues, of which one is terminal, in a histidine brace[82] (Fig 5.5). In some LPMO families, the Cu ion is also coordinated by a tyrosine residue. There exist multiple LPMO families (AA9-AA16) that catalyse the same reaction, but have slight differences in the amino acid sequence.

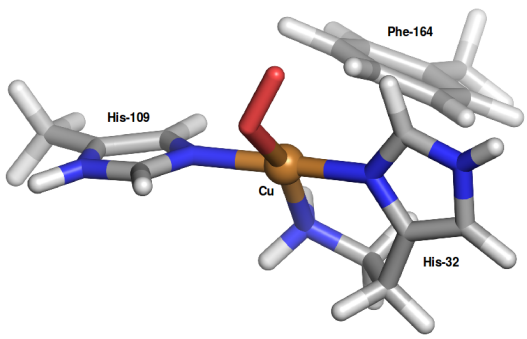


Figure 5.5 AA10-LPMO active site bound with an oxygen species bound to the Cu ion.

The X-ray structure of LPMOs was determined rather recently and, while several mechanisms have been proposed, its chemistry is still heavily under study. Issues arise even regarding the co-substrate.[83] Whereas molecular oxygen is the commonly accepted co-substrate of the LPMO reaction, it has recently been proposed that H_2O_2 could instead act as a co-substrate.[84] It is also known that LPMOs generate hydrogen peroxide in the absence of a substrate.[85]

Furthermore, the protonation of the active site is also not fully elucidated. Although there exist neutron structures of LPMO, which permit identification of hydrogen atoms, the protonation of the terminal N atom in the active site is still under debate.[86]

In paper V, we study the protonation of an AA1O-LPMO active site using neutron quantum refinement and the mechanism of hydrogen peroxide release by LPMOs using QM/MM calculations.

5.6 Triosephosphate isomerase

Triosephosphate isomerase (TIM) is an enzyme that catalyses the 1,2-proton shift of dihydroxyacetone phosphate (DHAP) to give (R)-glyceraldehyde 3-phosphate (GAP).[87] It has been observed that reducing the activity of the enzyme can cause severe diseases in humans.

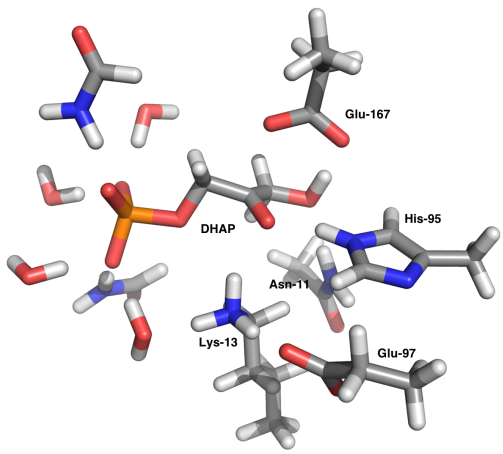


Figure 5.6 Triosephosphate isomerase active site with DHAP as substrate.

The active site of the enzyme contains four residues that form hydrogen bonds with the substrate, a glutamate, a histidine, an asparagine and a lysine[88] (Figure 5.6). It is believed that the asparagine and the lysine only provide electrostatic stabilisation.[89]

Two possible mechanisms for the isomerisation of DHAP have been proposed: In the classical mechanism, both the histidine and the glutamate residues are involved in proton transfer, whereas in the criss-cross mechanisms only the glutamate residue participates in proton transfer.[87]

To gain insight on the structure of the mechanistic intermediates of the TIM reaction, crystallographic studies of TIM inhibitors need to be performed. 2-phosphoglycolate (PGA) and phosphoglycolohydroxamate (PGH) have been proposed to mimic two reaction intermediates of the isomerisation.[90]

In paper VI, we use neutron quantum refinement to correctly identify the atomic position of the H atoms in the active site of two new neutron structures of TIM with the inhibitors described above.

6. Summary of the papers

The papers in this thesis all combine computational methods with crystallographic methods and data, but can be divided into several groups, depending on which computational method and which kind of crystallographic data was used:

- Papers I–VI use a combination of quantum chemical calculations and crystallographic refinement, called quantum refinement, to improve and understand the structure of the active site in several enzymes.
- Papers VII and VIII investigate what type of information about protein dynamics can be extracted directly from crystal structures.
- Papers IX and X attempt to develop methods to improve the modelling of ordered water molecules in both X-ray and neutron crystallographic data.

Paper I

In this paper we have studied the enzymatic active site of particulate methane monooxygenase (pMMO). As pMMO is a membrane protein, it has been very difficult to characterise experimentally from a structural point of view. For this reason, the metal content of the active site has been controversial, with proposals of dinuclear Fe centres as well as mono-, di- and trinuclear Cu active sites (Fig 6.1).[62–64]

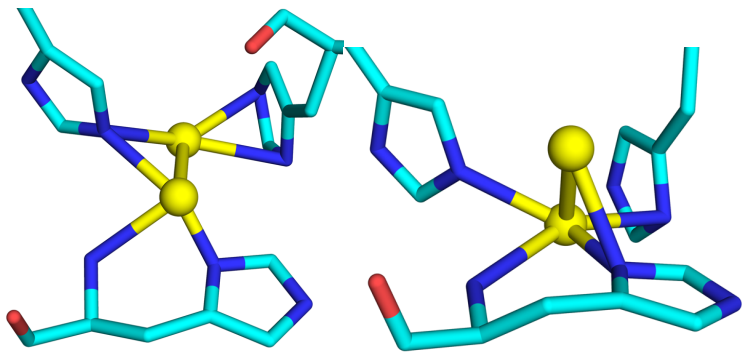


Figure 6.1 Examples of dinuclear copper sites in pMMO deposited structures.

We have studied two deposited crystal structures of pMMO, of rather poor resolution (2.8 and 2.68 Å respectively) with quantum refinement. We have tested several compositions and geometries of the active site that could fit the crystallographic raw data, including the original dinuclear copper site, as well as a modified dinuclear copper site, and a mononuclear copper site. We evaluated the resulting structures by their fit to the data through the RSZD scores but also through strain energies obtained by QM calculations (i.e. the energy difference of the active site when optimised in the crystal or in vacuum).

Electron-density maps from quantum refinement showed that the dinuclear site present in the deposited structures fits the crystallographic data rather poorly. In particular, one of the copper atoms gives rise to negative difference density, indicating that it is not correctly modelled. Modifying the geometry of the active site to a more chemically reasonable one with two additional bound water molecules did not improve the fit to the crystallographic data, as no electron density is seen around the water molecules.

However, the model with only one Cu atom fit the data well, both when it is bound only by the histidine brace or if there is an extra water molecule (Fig 6.2). Moreover, the mononuclear Cu sites gave the lowest strain energies, half as high as the deposited dinuclear Cu sites. Thus, we found no support for a dinuclear Cu site in the crystal structures of pMMO. Instead, the active site is better modelled with only one Cu atom.

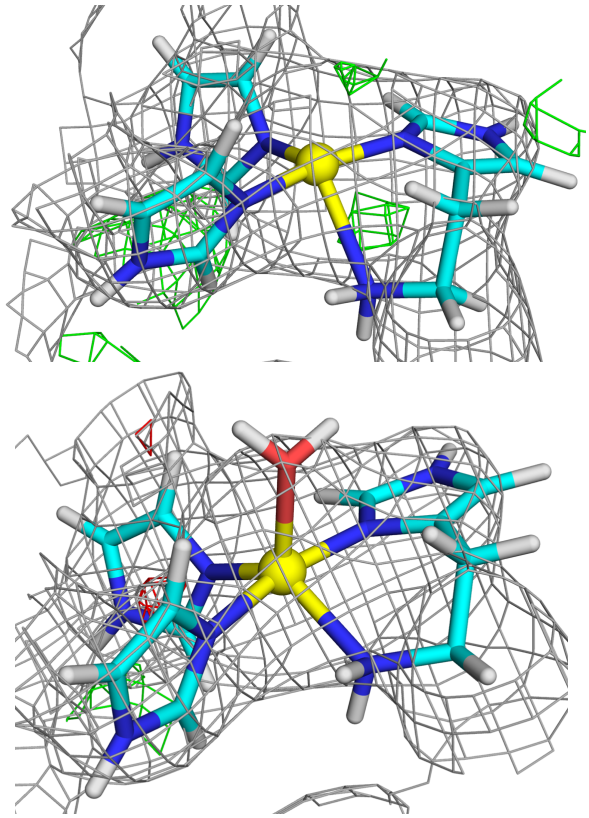


Figure 6.2 Quantum refined mononuclear copper active sites of pMMO without a water molecule (top) and with a water molecule coordinated to the copper atom (bottom).

Paper II

In this paper we studied the protonation states of the homocitrate ligand bound to the Mo ion in the active site of nitrogenase, as well as of other nearby residues. In solution, the pK_a values of homocitrate are known and a triply deprotonated state is most stable at neutral pH. However, this may change in the enzyme, as homocitrate is bound directly to a metal atom, which can change the electronic structure and the pK_a values of the ligand. The elucidation of the protonation states in nitrogenase is a prerequisite to perform reliable mechanistic studies of the enzyme.

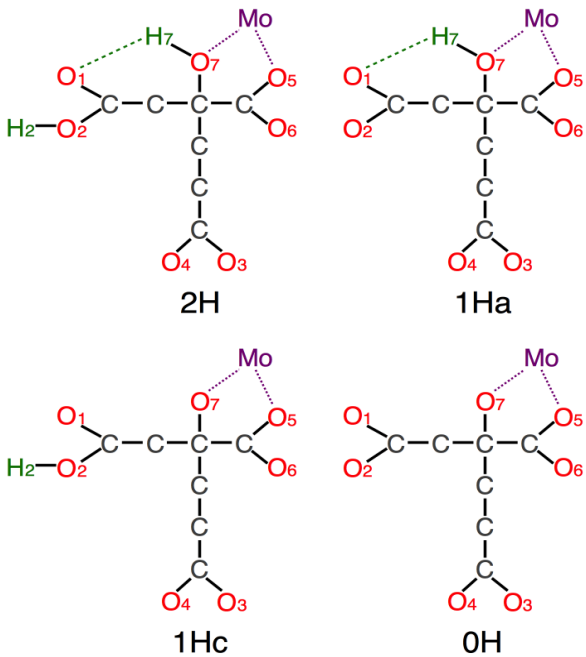


Figure 6.3 Protonation states of the homocitrate tested in paper II.

Although X-ray crystallography cannot directly identify hydrogen atoms, their presence is reflected in the position of the heavy atoms around them. As the QM/MM calculations include hydrogen atoms explicitly, quantum refinement can give good indications of protonation states.[91]

In paper II, quantum refinement was used as a complementary method for studying protonation, together with MD simulations, QM/MM and QM-cluster calculations. Only the homocitrate ligand protonation was investigated with quantum refinement.

We employed the 1.0 Å crystallographic structure and re-refined it replacing the E&H restraints with QM/MM restraints in four different protonation states of the homocitrate (Figure 6.3): a doubly deprotonated state, two triply deprotonated states and one fully deprotonated state. We evaluated the four refinements based on the maximum absolute RSZD score of the homocitrate ligand. The results show that there are vast differences between the RSZD of four states of the homocitrate, three of the states having a maximum RSZD higher than 3.0. The best state in quantum refinement is the triply deprotonated state, with the only protonation on the hydroxyl oxygen atom bound to molybdenum. This is also apparent in the electron density maps, which show that for the doubly deprotonated state, an oxygen atom is not in the correct position, giving rise to difference densities (Figure 6.4).

The protonation state concluded from quantum refinement is also supported by the QM/MM and QM-cluster calculations.

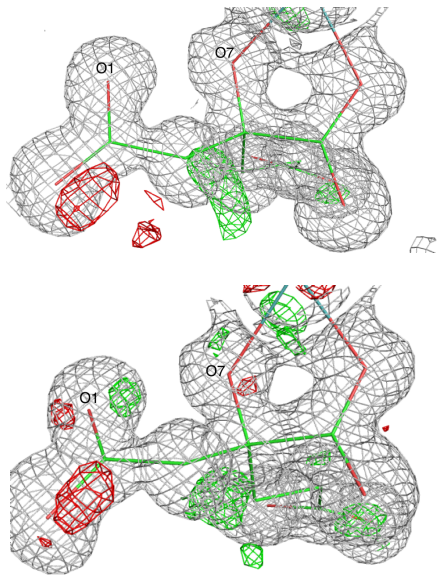


Figure 6.4 Electron-density maps of the 1Ha (top) and 2H (bottom) protonation states of the homocitrate ligand in the quantum refinement of the nitrogenase structure. The $2mF_o - DF_c$ maps are contoured at 1.0 σ and the $mF_o - DF_c$ maps are contoured at +3.0 σ (green) and −3.0 σ (red).

Paper III

The main purpose of paper III was to elucidate the reaction mechanism of sulfite oxidase. The mechanistic studies were performed through QM/MM calculations, including QM/MM free energy perturbations. Additionally, we studied what form of the molybdopterin ligand is involved in the enzyme. Three different forms of molybdopterin, with the phosphate group protonated (MPH), deprotonated (MPD) or in an oxidised (MPO) form could be possible from the data available in crystal structures (Figure 6.5). Traditional X-ray crystallography cannot discern between these forms, which differ only in the number of hydrogen atoms present. However, quantum refinement can be used to give indications of protonation states and is again used as a complementary technique to QM/MM calculations.

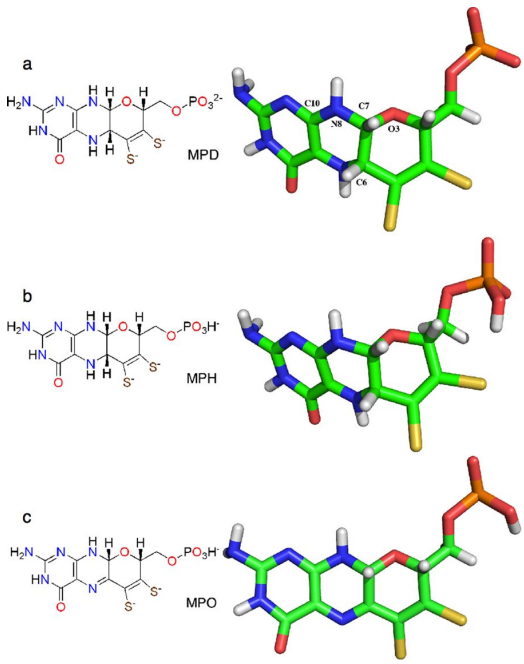


Figure 6.5 The three molybdopterin states studied in paper III: (a) deprotonated phosphate (b) protonated phosphate and (c) oxidised form.

The QM/MM calculations showed that only the S–OMo mechanism (see section 5.4 for nomenclature) is plausible, as the attempts to study the two other proposed

mechanisms failed. Regarding the nature of the molybdopterin ligand, the QM/MM calculations were less clear. Mechanisms with all three ligands showed reasonable reaction and activation energies, with slightly lower activation energies for the protonated and oxidised forms.

On the other hand, quantum refinement showed that the oxidised form, which has a more rigid geometry does not fit the electron density. The other two molybdopterin forms show a similar goodness of fit to the electron density from crystallographic data (Figure 6.6). A small positive difference density blob around the phosphate group of MPD and a higher total RSZD score for MPD indicate that the MPD group is slightly worse positioned in the electron density. This suggests that MPH is the best model of molybdopterin in sulfate oxidase, with the phosphate group protonated. This conclusion was also strengthened by hydrogen bond analysis in the QM/MM structures.

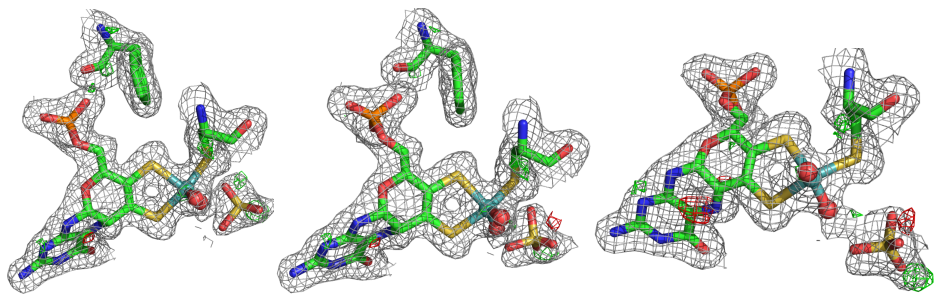


Figure 6.6 Results of the quantum refinements of sulfite oxidase with MPD (left), MPH (middle) and MPO (right). The $2mF_o - DF_c$ maps are contoured at 1.0 σ (gray) and the $mF_o - DF_c$ maps are contoured at +3.0 σ (green) and −3.0 σ (red).

Paper IV

Paper IV deals with development and testing of quantum refinement for protein neutron crystal structures. Improved restraints are especially important in neutron structures, as the statistical geometrical information available for bonds involving H atoms is poorer than for heavier atoms, making the refined positions of H atoms less reliable. This involves replacing the E&H restraints in joint X-ray/neutron refinement with QM/MM restraints, analogous to the way quantum refinement works in X-ray crystallography.

The method is implemented as an interface between the *nCNS* crystallography software, capable of running joint X-ray/neutron refinement and the *Turbomole* software for quantum chemical calculations. In the paper, we apply neutron quantum refinement to lactose in the binding site of galectin-3C (Figure 6.7).

We show that the method behaves properly and that it can improve water orientations and hydrogen atom positions in the ligand compared to both traditional joint refinement and to pure QM/MM minimisations. In particular, the hydrogen-bonding pattern in the quantum system is greatly improved. Additionally, we show that we can vary the weights of the neutron data, X-ray data or QM/MM restraints in order to get an optimal structure and that RSZD values of atoms in the quantum system give a good validation of the weights.

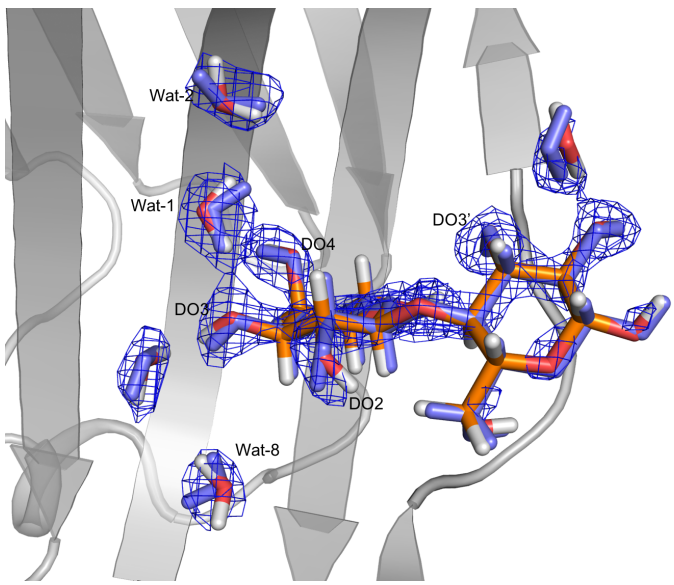


Figure 6.7 Overlay of the lactose-galectin-3 deposited structure before (blue) and after QM refinement (carbon atoms in orange). The nuclear $2m|F_o| - D|F_c|$ density at 0.8 σ is shown in blue.

We also apply the method to a metalloenzyme for which neutron data exists, lytic polysaccharide monooxygenase (LPMO). Results show that neutron quantum refinement is also useful for discerning between protonation states. This is studied in more detail in paper V.

Paper V

The purpose of this paper was to study the mechanism of hydrogen peroxide generation by an LPMO belonging to the AA10 family. In order to perform correct QM/MM mechanistic calculations, we first elucidated the protonation and oxidation state of residues in the active site, including the copper atom and the oxygen moiety, based on recent X-ray and neutron structures.

A previous LPMO neutron structure with an oxygen species bound suggested a deprotonated N-terminus on the histidine binding to the Cu in one of the two subunits of the protein and a protonated N-terminus in the other. We performed joint X-ray/neutron refinement and joint X-ray/neutron quantum refinement of this structure to determine the protonation state of the N-terminus. The structures with only one D atom on the N-terminal atom gave rise to positive difference density around the N-terminus in both subunits, albeit slightly lower in one subunit, with both traditional refinement and quantum refinement. Thus, we concluded that the N-terminus is not deprotonated in the LPMO active site (Figure 6.8).

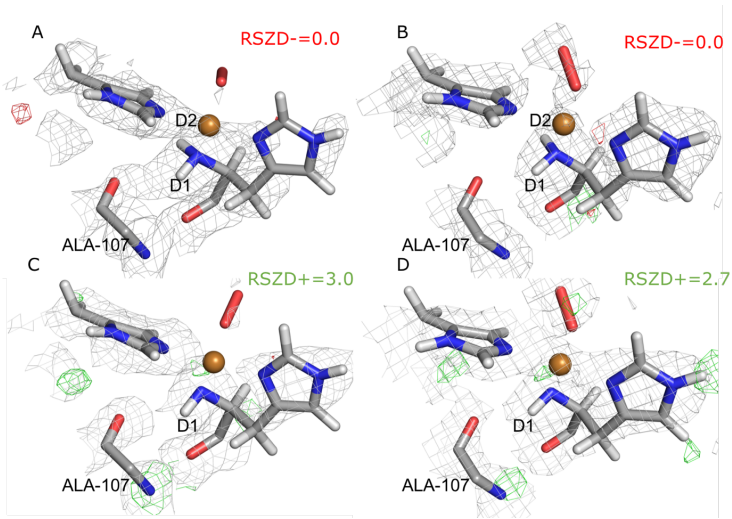


Figure 6.8 Structure and nuclear density maps of the LPMO active site after quantum refinement. $m2F_O - DF_C$ maps are contoured at 1.0σ and $mF_O - DF_C$ maps are contoured at +2.8σ (green) and −2.8σ (red) (A) – subunit A, ND_2; (B) – subunit B, ND_2; (C) – subunit A, ND^-, (D) – subunit B, ND^-.

Interestingly, the main difference between the join quantum refinement and traditional joint refinement was in the geometry of the oxygen species, which suggested that its nature might have not been assigned correctly in the crystal structure. Consequently, we performed another round of quantum refinement with different oxygen species, including both $[CuO_2]$ and $[CuO_2]^+$. For this, only the X-ray data was used, as the nuclear density around the oxygen atoms was rather weak. RSZD values and strain energies calculated by quantum refinement showed that $[CuO_2]^+$ fits better to the X-ray data in both subunits, suggesting a superoxide nature of the oxygen species (Figure 6.9).

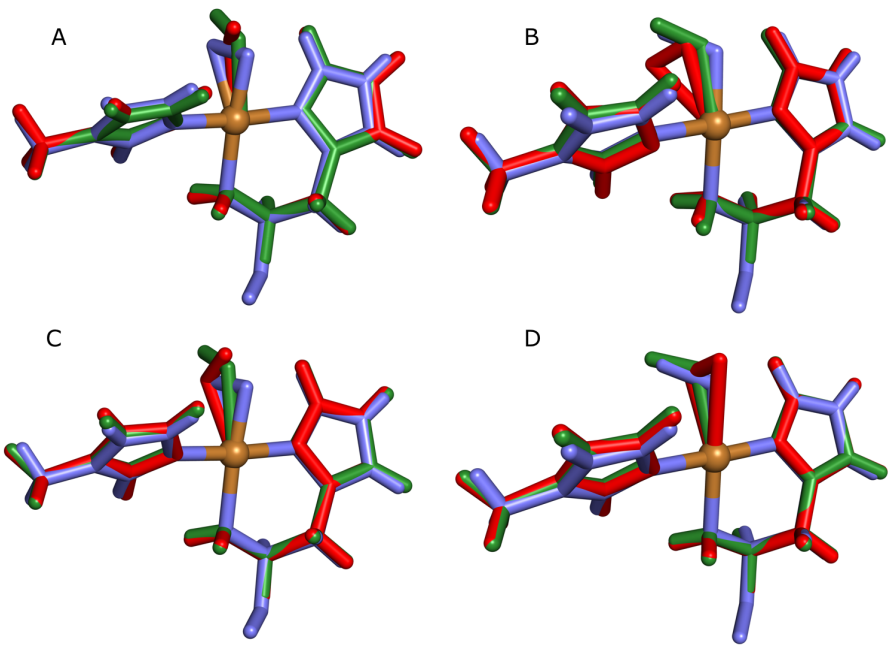


Figure 6.9 LPMO active sites in the original crystal structure (blue), calculated/quantum refined structures with peroxide oxygen species (yellow) or superoxide oxygen species (red). A – QM/MM structure, subunit A; B – QM/MM structure, subunit B; C – quantum refined structure, subunit A; D – quantum refined structure, subunit B.

Then, we ran QM/MM calculation to investigate the hydrogen peroxide generation mechanism, starting from the best structures obtained in quantum refinement. The results showed that in the most energetically favourable mechanism, hydrogen peroxide forms at the copper centre before dissociating and

the copper must be in the Cu(I) when it dissociates. The consecutive protonations occur through a nearby glutamate residue.

Paper VI

In this paper, neutron crystal structures of TIM from *Leishmania mexicana* with two different inhibitors, 2-phosphoglycolate (PGA) and phosphoglycolohydroxamate (PGH)were studied. Both traditional joint X-ray/neutron refinement and neutron quantum refinement were performed to determine the correct positions of protons on the inhibitors and in the protein active site. Of particular interest were protonation states of residues that are assumed to be involved in the isomerization reaction mechanism, i.e. Glu-167 and His-95, but also residues that contribute to the electrostatic stabilization of the substrate, i.e. Lys-13 and Glu-87. This paper gave the first experimental evidence for the protonation state of the active site, from neutron diffraction experiments.

Refinement of the PGA-TIM complex showed that Glu-167 is protonated and His-85 is protonated on the Nε atom, forming a bifurcate hydrogen bond with the inhibitor (Figure 6.10). Additionally, the nuclear density maps suggested that Lys-13 is positively charged, whereas the inhibitor is deprotonated and negatively charged. All possible protonation states were tested by quantum refinement, which confirmed the initial proton assignment as the most energetically stable state that also fits the nuclear density maps.

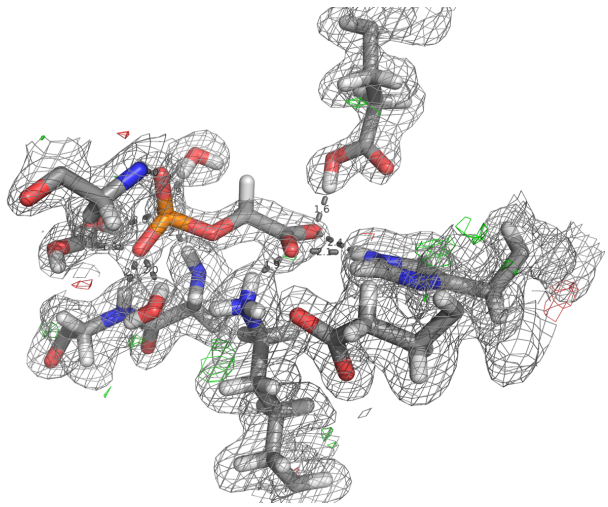


Figure 6.10 Joint X-ray—neutron refined neutron structure of PGA-TIM with the $2mF_o$-DF_c nuclear scattering length density maps contoured at 1.0 σ and the mF_o–DF_c nuclear scattering length difference density maps contoured at 3.0 σ (green) and –3.0 σ (red).

The neutron structure of PGH-TIM was less conclusive and gave little information on the protonation states of the inhibitor and of Glu-167. However, it wass clear that His-95 is also protonated on the Nε atom in this structure. Quantum refinement calculations suggested that the PGH inhibitor is protonated, whereas the Glu-167 residue is deprotonated. Unfortunately, this is less evident from the nuclear density maps (Figure 6.11), possibly because of disorder or of negative interference between the aliphatic H atoms of the inhibitor and the exchangeable D atoms.

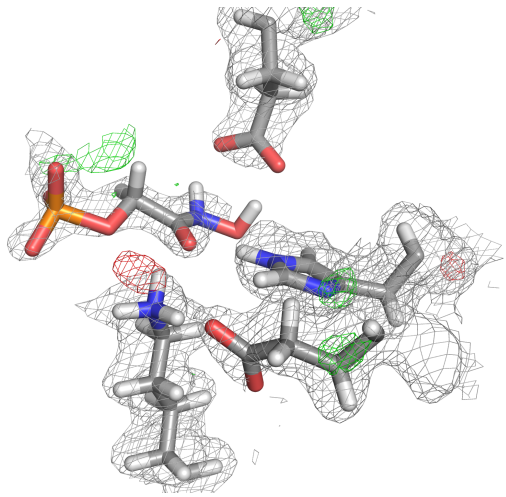


Figure 6.11 Quantum refined neutron structure of PGH-TIM with the $2mF_o$-DF_c nuclear scattering length density maps contoured at 1.0 σ and the mF_o–DF_c nuclear scattering length difference density maps contoured at 3.0 σ (green) and –3.0 σ (red).

QM/MM calculations showed that the natural substrate of TIM, DHAP, fits well in the active site geometry found in the PGA-TIM neutron structure. Furthermore, the first intermediate in the reaction mechanism, of which PGA-TIM is a mimic, has a very similar geometry to the corresponding neutron structure.

A full QM/MM calculation of both proposed mechanisms was also performed starting from the PGA-TIM neutron structure, suggesting that the criss-cross mechanism is energetically more reasonable than the classical mechanism.

Paper VII

In this paper, we used ensemble refinement to study ligand dynamics in complexes of galectin-3C with several ligands, including the natural ligand lactose.

Ensemble refinement simulations of six protein-ligand complexes revealed a large amount of flexibility of the ligands and some residues in the binding site of galectin-3C, much larger than what can be observed in traditional. crystallographic refinement. For some ligands, it showed up to 100 different possible conformations in the active site (Figure 6.12).

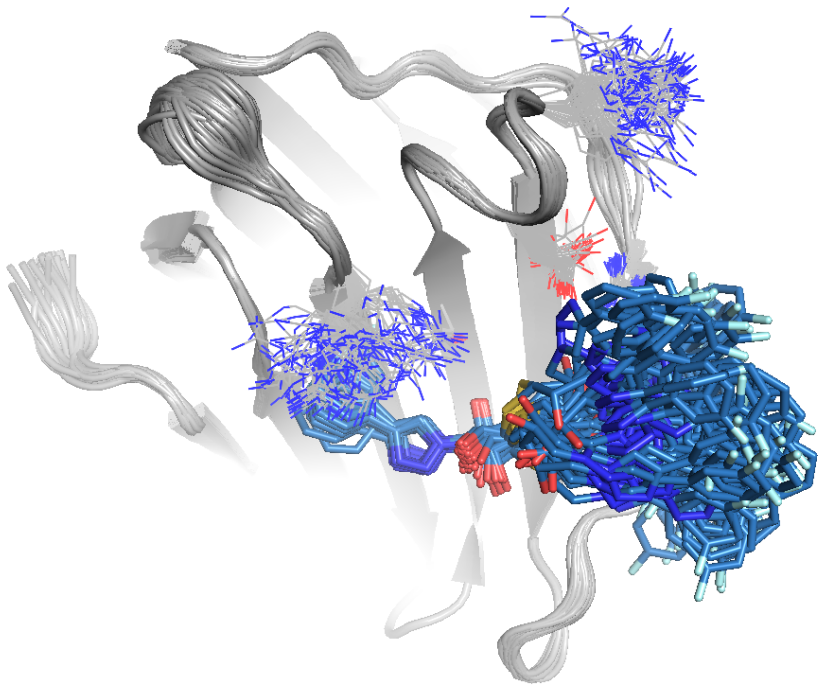


Figure 6.12 Ensemble of structures resulting from ensemble refinement of S-galectin-3C.

We also compared the results of ensemble refinement with other computational methods to study dynamics and generate alternative conformations, such as MD simulation or *qFit-ligand*. The MD simulations also showed a large amount of flexibility for most ligands, confirming the observations of ensemble refinement. *qFit-ligand* modeled four out of the six ligands in multiple alternative

conformations and the remaining ligands were those that showed the least amount of flexibility in ensemble refinement.

Ensemble refinement thus gave a significantly different view of the bound ligand structure compared to traditional refinement. While the results should not be trusted quantitively, ensemble refinement gave good indication of what parts of the crystal structure were reliable and which may involve significant dynamics.

Paper VIII

In this paper, we attempted to calculate the protein conformational entropy starting from crystallographic B-factors. As test cases we have used both cryo and room-temperature structures of galectin-3C in complex with two diastereomeric ligands, as well as high resolution cryo structures of lysozyme in complex with two closely related ligands and trypsin in complex with two similar ligands. Only relative conformational entropies were calculated, such that error cancellation may occur. All structures were re-refined in order to maximise the agreement between the alternative conformations present in each of the two complexes and minimise differences caused by variations in the refinement procedures of different crystallographers.

The entropy calculated from B-factors was compared to experimental data from nuclear magnetic resonance (NMR) and isothermic titration calorimetry (ITC) for the galectin-3C complexes, as well as with more traditional methods of calculating conformational entropy, such as dihedral histogramming or quasi-harmonic analysis (see section 3.4 for a more detailed description of these methods). The method for calculating entropy directly from B-factor was first proposed by Zagrovic et al[92] and is also based on quasi-harmonic analysis. The B-factors used for the entropy calculation were either isotropic or anisotropic and they were obtained either from a normal refinement or with a TLS model of the whole protein.

The results showed that isotropic B-factors greatly overestimate the conformational entropy, giving results 40 times larger than the experimental total entropy or than the calculated conformational entropy from MD simulations. The use of anisotropic B-factors slightly lowered the conformational entropy but it was still too high to be usable. Using a TLS model brought the entropy to the same order of magnitude as in the MD simulations, but gave the incorrect sign. Using B-factors from the room temperature structures of the galectin-3C complexes improved the entropy estimates, as these B-factors reflect more accurately the real movements inside the protein. Unfortunately, the resulting entropy was still ~5 times larger than that obtained from MD simulations and TLS refinement did not reduce the entropy further for room temperature structures.

We also ran MD simulation in a crystallographic unit cell, to test if the difference we observe between B-factors and MD simulation stem from the different movements in solution and crystal. However, the entropy obtained from crystal MD simulations was very similar to the one calculated from traditional MD simulations. Next, we attempted to include some of the correlation in the entropy calculation from B-factors, by including covariance terms from MD simulations

in the variance-covariance matrix used in entropy calculation. This gave very poor results, suggesting that B-factors from crystallographic refinement are not compatible with B-factors from MD simulations.

Thus, we had to conclude that it is currently not possible to extract meaningful entropies from B-factors, even after careful re-refinement or using room-temperature data.

Paper IX

This paper presents a comparison of the water structure between protein crystal structures and MD simulations in solution or in the crystal. To this end, we performed MD simulations both in solution and in a crystallographic unit cell of galectin-3C in complex with two ligands. These simulations were compared to the corresponding crystal structures collected at both cryogenic temperature and at room temperature.

Analysis of the water structures in the MD simulations was performed in two ways. Firstly, the water-molecule density was calculated using global, grid-based clustering and peaks were identified in the resulting density maps. These peaks were then compared to the positions of crystallographic water molecules by measuring the minimum distance between each crystal water and a peak in the MD simulation. The results for the grid-based analysis showed that the MD simulations could only partly reproduce the water structure found in the crystal structures. MD performed in crystal showed better results than solution MD, which was expected, as the environment in crystal MD is more similar to the one in a crystal structure. However, only a maximum of 31% of crystal-water molecules in the cryo-structure had a neighbour within 1 Å in the MD simulations.

Secondly, crystallographic water sites were defined based on the distances between the water molecule in the crystal structure and heavy atoms of the protein. Then, the water sites were followed throughout the MD simulations and the closest MD water molecules were recorded and clustered. This local clustering analysis yielded much better results, with a maximum of 82% of water sites reproduced by the MD simulations. These results suggest that MD simulations are able to reproduce the distribution of water molecules around the protein that is observed in crystal structures. However, protein dynamics need to be correctly taken into account when analysing the MD simulations.

We also attempted to insert new water molecules in the crystal structures if there was experimental evidence for them in the electron density maps. Crystal MD of one of the ligands found water molecules that were close to unmodeled density regions in the crystallographic maps. Three new water molecules could be inserted into the cryo-structure and into the room temperature structure of one of the complexes.

Paper X

In this paper, we developed a method to automatically decide the positions of water deuterium atoms in neutron structures. To this end, we first tested three validation metrics for the evaluation of deuterium positions in neutron structures. This was performed by a qualitative evaluation of each water molecule in ten different structures and then comparing this evaluation to B-factors, RSZD and RSCC values. The RSCC performed the best among the three, yielding only 22% false positives or negatives. We found that a RSCC threshold of 0.81 was appropriate to decide whether a water orientation was good or poor.

We also performed multiple refinements of the neutron structure of galectin-3C in order to find a refinement strategy that resulted in the highest number of water molecules in good orientations. The best results were obtained with 15 cycles of refinement in reciprocal space using only neutron data and a fixed protein.

Finally, a script was developed to automatically generate water orientations. The script first calculates the RSCC values for all water molecules in the current structure. It then generates a new orientation for all water molecules with at least one deuterium atom with a RSCC value below the threshold. This is repeated until all water molecules in the structure are in good orientations according to their RSCC values or until a maximum number of orientations has been tested. Optionally, a crystallographic refinement was conducted for each new orientation.

This procedure was applied to three protein neutron structures. The results showed that the number of properly oriented water molecules increased in all proteins after using the script and refining the final structure compared to only performing the refinement. Moreover, the number of water molecules in good orientations after running the automated reorientation script was also higher than in the deposited structures. Running the reorientation script without a refinement after each cycle did not take longer than the refinement itself for two out of the three proteins studied, so it could be used alongside refinement for any protein structure without a significant increase in computing time.

7. Conclusions and Outlook

The purpose of this thesis has been to develop and apply some methods that combine computational chemistry and macromolecular crystallography. These methods improve the quality of the protein structure obtained from X-ray and neutron diffraction and the quality and accuracy of computational studies, by incorporating experimental information.

Firstly, I have extensively used a previously developed method which combines quantum-mechanical calculations with crystallographic refinement, called quantum refinement.[46,47] In papers I–III, I show the usefulness of quantum refinement in determining the structures of metal sites in some metalloenzymes. For low- and intermediate-resolution structures, such as the pMMO structure shown in paper I, quantum refinement reveals huge improvements in the metal-site geometry and we could draw conclusions about its composition with much higher confidence than from traditional crystallographic refinement. Quantum refinement also proved useful for higher resolution structures, such as nitrogenase studied in paper II and sulfite oxidase studied in paper III. While the position of heavy atoms is adequately described by traditional methods, we were able to predict the protonation states of various groups in the active sites of the metalloenzymes. However, quantum refinement of ultra-high resolution (<1.0 Å) failed (because then systematic errors of the QM methods start to be significant) and such studies are not included in the thesis. Other ways to include quantum mechanical information in ultra-high resolution structures of proteins are currently developed in our group, by improving the structure factors directly from the density obtained from QM calculations, through Hirshfeld-atom refinement.[93]

In this thesis, I have also extended quantum refinement for treating neutron structures. Papers IV–VI show that neutron quantum refinement works and can confirm and improve protonation states of ligands and protein residues or water molecule orientations found in traditional joint X-ray/neutron refinement. In particular, the hydrogen-bonding pattern present in protein active sites is improved in neutron quantum refinement.

Both X-ray and neutron quantum refinement are currently implemented only in a rather old crystallographic software, CNS. Thus, they do not take advantage of all modern crystallographic methods. An implementation in the Phenix software is currently underway and should improve the results of quantum refinement.

This would allow for better bulk-solvent treatment, B-factor refinement and eventually the use of force fields other than E&H for the parts not included in the quantum system.

Another subject in this thesis is how to obtain, treat and describe dynamics information in crystal structures . Papers VII and VIII show that this is a complex subject and it is not yet clear how to best use the information on atomic movements that we derive from diffraction experiments. Paper VII shows an alternative method to express dynamics in crystal, through ensemble refinement. The study reveals that parts of protein on the surface possess a much larger flexibility than is indicated by alternative conformations and B-factors. This is especially important for ligands bound to the surface of the protein and could help us to understand ligand binding better. Even though ensemble refinement is not a fully quantitative method, we hope that it is used more widely in the future to provide insight into more flexible groups in the proteins and the ligands bound to them.

In paper VIII I attempt to use B-factors to calculate protein conformational entropy. Results show that this is not yet possible and B-factors do not correlate well with movements we observe in MD simulations. Although the results are quite disappointing, the information provided by B-factors, alternative conformations or ensembles from ensemble refinement is still valuable and should be included in more computational studies. The exact way to do this is still elusive but could be assessed in further research.

Finally, the last two papers in this thesis deal with water molecules in X-ray and neutron protein crystal structures, respectively. In paper IX, we show that MD simulations are able to reproduce the crystallographic water structure, but a correct definition of water sites and subsequent local clustering needs to be performed. In contrast, a traditional global clustering of water molecules gives rather poor results. Furthermore, MD simulations could be used to provide evidence for new water molecules that are not well-ordered. With the advent of GPUs, making 100 ns of MD simulation time routine, it could be useful to include MD simulations in crystal in more crystallographic studies.

Paper X presents a method for automatically finding good orientations of water molecules in neutron structures. This method shows significant improvements compared to only doing refinement for all proteins tested. However, it could be further improved by finding a better validation metric for what constitutes a good orientation of a water molecule in a nuclear density map. Machine learning techniques may be the answer for an ideal validation metric.

In conclusion, bringing macromolecular crystallography and computational chemistry closer together is possible and can improve both fields. However, many

more studies are necessary before we can unequivocally know when and how to use different potentials in refinement or how to best express and use dynamics in crystal structures.

References

(1) Moriarty, N. W.; Janowski, P. A.; Swails, J. M.; Nguyen, H.; Richardson, J. S.; Case, D. A.; Adams, P. D. Improved Chemistry Restraints for Crystallographic Refinement by Integrating the Amber Force Field into Phenix. *bioRxiv* **2019**, 724567.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(3) Kendrew, J. .; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181* (4610), 662–666.

(4) Ducruix, A.; Giegé, R.. *Crystallization of Nucleic Acids and Proteins : A Practical Approach*; Oxford University Press, 1999.

(5) Karplus, P. A.; Diederichs, K. Assessing and Maximizing Data Quality in Macromolecular Crystallography. *Curr. Opin. Struct. Biol.* **2015**, *34*, 60–68.

(6) Karplus, P. A.; Diederichs, K. Linking Crystallographic Model and Data Quality. *Science (80-.).* **2012**, *336* (6084), 1030–1033.

(7) Usón, I.; Sheldrick, G. M. An Introduction to Experimental Phasing of Macromolecules Illustrated by SHELX; New Autotracing Features. *Acta Crystallogr. Sect. D, Struct. Biol.* **2018**, *74* (Pt 2), 106–116.

(8) Keedy, D. A.; Fraser, J. S.; van den Bedem, H. Exposing Hidden Alternative Backbone Conformations in X-Ray Crystallography Using QFit. *PLOS Comput. Biol.* **2015**, *11* (10), e1004507.

(9) Urzhumtsev, A.; Afonine, P. V; Adams, P. D. TLS from Fundamentals to Practice. *Crystallogr. Rev.* **2013**, *19* (4), 230–270.

(10) Engh, R. A.; Huber, R. Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement. *Acta Crystallogr. Sect. A* **1991**, *47* (4), 392–400.

(11) Tronrud, D. E.; Karplus, P. A. A Conformation-Dependent Stereochemical Library Improves Crystallographic Refinement Even at Atomic Resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67* (Pt 8), 699.

(12) Jiang, J.-S.; Brünger, A. T. Protein Hydration Observed by X-Ray Diffraction. *J. Mol. Biol.* **1994**, *243* (1), 100–115.

(13) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; et al. *PHENIX* : A Comprehensive Python-Based System for Macromolecular

Structure Solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(14) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. Sect. D* **1997**, *53*, 240–255.

(15) Smart, O. S.; Womack, T. O.; Flensburg, C.; Keller, P.; Paciorek, W.; Sharff, A.; Vonrhein, C.; Bricogne, G. Exploiting Structure Similarity in Refinement: Automated NCS and Target-Structure Restraints in *BUSTER*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2012**, *68* (4), 368–380.

(16) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963**, *7* (1), 95–99.

(17) Atkins, P. W.; Friedman, R. *Molecular Quantum Mechanics*; Oxford University Press, 2011.

(18) Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.* **1928**, *24* (1), 89–110.

(19) Huzinaga, S. Gaussian-Type Functions for Polyatomic Systems. I. *J. Chem. Phys.* **1965**, *42* (4), 1293–1302.

(20) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley, 2001.

(21) Jensen, F. *Introduction to Computational Chemistry*, 3rd ed.; John Wiley & Sons, Ltd: Chichester, 2017.

(22) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133–A1138.

(23) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(24) Becke, A. D. Density-Functional Exchange-Energy Approximation With Correct Asymptotic-Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

(25) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Non-Empirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91* (14), 146401.

(26) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

(27) Becke, A. D. A New Mixing of Hartree–Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98* (2), 1372.

(28) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104 (19 pages).

(29) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 383–394.

(30) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(31) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676.

(32) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.

(33) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103* (2), 227–249.

(34) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A New ONIOM Implementation in Gaussian98. Part I. The Calculation of Energies, Gradients, Vibrational Frequencies and Electric Field Derivatives. *J. Mol. Struct. THEOCHEM* **1999**, *461–462*, 1–21.

(35) Assfeld, X.; Rivail, J.-L. Quantum Chemical Computations on Parts of Large Molecules: The Ab Initio Local Self Consistent Field Method. *Chem. Phys. Lett.* **1996**, *263* (1–2), 100–106.

(36) Jiali Gao, *,†; Patricia Amara, ‡; Cristobal Alhambra, † and; Martin J. Field*, ‡. A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations. **1998**.

(37) Ryde, U. The Coordination of the Catalytic Zinc in Alcohol Dehydrogenase Studied by Combined Quantum-Chemical and Molecular Mechanics Calculations. *J. Comput. Aided. Mol. Des.* **1996**, *10*, 153–164.

(38) Ryde, U.; Olsson, M. H. M. Structure, Strain, and Reorganization Energy of Blue Copper Models in the Protein. *Int. J. Quantum Chem.* **2001**, *81*, 335–347.

(39) Kr□utler, V.; van Gunsteren, W. F.; H□nenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508.

(40) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998**, *282* (5389), 740–744.

(41) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(43) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in

Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–6271.

(44) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational Entropy Changes upon Lactose Binding to the Carbohydrate Recognition Domain of Galectin-3. *J. Biomol. NMR* **2009**, *45* (1–2), 157–169.

(45) Case, D. A. Normal Mode Analysis of Protein Dynamics. *Curr. Opin. Struct. Biol.* **1994**, *4* (2), 285–290.

(46) Ryde, U.; Nilsson, K. Quantum Chemistry Can Locally Improve Protein Crystal Structures. *J. Am. Chem. Soc.* **2003**, *125* (47), 14232–14233.

(47) Ryde, U.; Olsen, L.; Nilsson, K. Quantum Chemical Geometry Optimizations in Proteins Using Crystallographic Raw Data. *J. Comput. Chem.* **2002**, *23* (11), 1058–1070.

(48) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (2), 91–100.

(49) Brunger, A. T. Version 1.2 of the Crystallography and NMR System. *Nat. Protoc.* **2007**, *2* (11), 2728–2733.

(50) Borbulevych, O. Y.; Plumley, J. A.; Martin, R. I.; Merz, K. M.; Westerhoff, L. M. Accurate Macromolecular Crystallographic Refinement: Incorporation of the Linear Scaling, Semiempirical Quantum-Mechanics Program DivCon into the PHENIX Refinement Package. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2014**, *70* (5), 1233–1247.

(51) Borbulevych, O.; Martin, R. I.; Tickle, I. J.; Westerhoff, L. M. XModeScore: A Novel Method for Accurate Protonation/Tautomer-State Determination Using Quantum-Mechanically Driven Macromolecular X-Ray Crystallographic Refinement. *Acta Crystallogr. Sect. D* **2016**, *72* (4), 586–598.

(52) Zheng, M.; Reimers, J. R.; Waller, P.; Afonine, P. V. Q | R : Quantum-Based Refinement Research Papers. **2017**, 45–52.

(53) Burnley, B. T.; Afonine, P. V; Adams, P. D.; Gros, P. Modelling Dynamics in Protein Crystal Structures by Ensemble Refinement. **2012**, 1–29.

(54) Gros, P.; van Gunsteren, W. F.; Hol, W. G. Inclusion of Thermal Motion in Crystallographic Structures by Restrained Molecular Dynamics. *Science* **1990**, *249* (4973), 1149–1152.

(55) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (June), 3684–3690.

(56) Forneris, F.; Burnley, B. T.; Gros, P. Ensemble Refinement Shows Conformational Flexibility in Crystal Structures of Human Complement Factor D. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2014**, *70* (3), 733–743.

(57) Dumic, J.; Dabelic, S.; Flögel, M. Galectin-3: An Open-Ended Story. *Biochim. Biophys. Acta - Gen. Subj.* **2006**, *1760* (4), 616–635.

(58) Leffler, H.; Carlsson, S.; Hedlund, M.; Qian, Y.; Poirier, F. Introduction to Galectins. *Glycoconj. J.* **2002**, *19*, 433–440.

(59) Liu, F. T.; Rabinovich, G. A. Galectins: Regulators of Acute and Chronic Inflammation. *Ann. N. Y. Acad. Sci.* **2010**, *1183*, 158–182.

(60) Sazinsky, M. H.; Lippard, S. J. *Methane Monooxygenase: Functionalizing Methane at Iron and Copper*; 2015; Vol. 15.

(61) Culpepper, M. A.; Rosenzweig, A. C. Architecture and Active Site of Particulate Methane Monooxygenase. *Crit. Rev. Biochem. Mol. Biol.* **2012**, *47* (6), 483–492.

(62) Lieberman, R. L.; Shrestha, D. B.; Doan, P. E.; Hoffman, B. M.; Stemmler, T. L.; Rosenzweig, A. C. Purified Particulate Methane Monooxygenase from Methylococcus Capsulatus (Bath) Is a Dimer with Both Mononuclear Copper and a Copper-Containing Cluster. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 3820–3825.

(63) Wang, V. C.-C.; Maji, S.; Chen, P. P.-Y.; Lee, H. K.; Yu, S. S.-F.; Chan, S. I. Alkane Oxidation: Methane Monooxygenases, Related Enzymes, and Their Biomimetics. *Chem. Rev.* **2017**, *117*, 8574–8621.

(64) Martinho, M.; Choi, D. W.; DiSpirito, A. A.; Antholine, W. E.; Semrau, J. D.; Münck, E. Mössbauer Studies of the Membrane-Associated Methane Monooxygenase from Methylococcus Capsulatus Bath: Evidence for a Diiron Center. *J. Am. Chem. Soc.* **2007**, *129* (51), 15783–15785.

(65) Hoffman, B. M.; Lukoyanov, D.; Yang, Z.-Y.; Dean, D. R.; Seefeldt, L. C. Mechanism of Nitrogen Fixation by Nitrogenase: The Next Stage. *Chem. Rev.* **2014**, *114* (8), 4041–4062.

(66) Spatzal, T.; Aksoyoglu, M.; Zhang, L.; Andrade, S. L. a.; Schleicher, E.; Weber, S.; Rees, D. C.; Einsle, O. Evidence for Interstitial Carbon in Nitrogenase FeMo Cofactor. *Science (80-.).* **2011**, *334* (November), 940–940.

(67) Eady, R. R. Structure−Function Relationships of Alternative Nitrogenases. *Chem. Rev.* **1996**, *96* (7), 3013–3030.

(68) Dance, I. Activation of N2, the Enzymatic Way. *Zeitschrift für Anorg. und Allg. Chemie* **2015**, *641*, 91–99.

(69) Varley, J. B.; Wang, Y.; Chan, K.; Studt, F.; Nørskov, J. K. Mechanistic Insights into Nitrogen Fixation by Nitrogenase Enzymes. *Phys. Chem. Chem. Phys.* **2015**, *17* (44), 29541–29547.

(70) Siegbahn, P. E. M.; Westerberg, J.; Svensson, M.; Crabtree, R. H. Nitrogen Fixation by Nitrogenases: A Quantum Chemical Study. *J. Phys. Chem. B* **1998**, *102* (9), 1615–1623.

(71) Feng, C.; Tollin, G.; Enemark, J. H. Sulfite Oxidizing Enzymes. *Biochim. Biophys. Acta - Proteins Proteomics* **2007**, *1774* (5), 527–539.

(72) Kappler, U.; Enemark, J. H. Sulfite-Oxidizing Enzymes. *J. Biol. Inorg. Chem.* **2015**, *20* (2), 253–264.

(73) Schrader, N.; Fischer, K.; Theis, K.; Mendel, R. R.; Schwarz, G.; Kisker, C. The Crystal Structure of Plant Sulfite Oxidase Provides Insights into Sulfite Oxidation in Plants and Animals. *Structure* **2003**, *11* (10), 1251–1263.

(74) Hille, R. Mechanistic Aspects of the Mononuclear Molybdenum Enzymes. *J. Biol. Inorg. Chem.* **1997**, *2* (6), 804–809.

(75) Das, S. K.; Chaudhury, P. K.; Biswas, D.; Sarkar, S. Modeling for the Active Site of Sulfite Oxidase: Synthesis, Characterization, and Reactivty of [MoVIO2(Mnt)2]2- (Mnt2- = 1,2-Dicyanoethylenedithiolate). *J. Am. Chem. Soc.* **1994**, *116* (1), 9061–9070.

(76) Thapper, A.; Deeth, R. J.; Nordlander, E. Computer Modeling of the Oxygen-Atom Transfer Reaction between Hydrogen Sulfite and a Molybdenum(VI) Dioxo Complex. *Inorg. Chem.* **1999**, *38* (5), 1015–1018.

(77) Rothery, R. A.; Stein, B.; Solomonson, M.; Kirk, M. L.; Weiner, J. H. Pyranopterin Conformation Defines the Function of Molybdenum and Tungsten Enzymes. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (37), 14773–14778.

(78) Harris, P. V.; Welner, D.; McFarland, K. C.; Re, E.; Navarro Poulsen, J.-C.; Brown, K.; Salbo, R.; Ding, H.; Vlasenko, E.; Merino, S.; et al. Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family. *Biochemistry* **2010**, *49* (15), 3305–3316.

(79) Vaaje-Kolstad, G.; Westereng, B.; Horn, S. J.; Liu, Z.; Zhai, H.; Sørlie, M.; Eijsink, V. G. H. An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides. *Science* **2010**, *330* (6001), 219–222.

(80) Meier, K. K.; Jones, S. M.; Kaper, T.; Hansson, H.; Koetsier, M. J.; Karkehabadi, S.; Solomon, E. I.; Sandgren, M.; Kelemen, B. Oxygen Activation by Cu LPMOs in Recalcitrant Carbohydrate Polysaccharide Conversion to Monomer Sugars. *Chem. Rev.* **2018**, *118* (5), 2593–2635.

(81) Tandrup, T.; Frandsen, K. E. H.; Johansen, K. S.; Berrin, J.-G.; Lo Leggio, L. Recent Insights into Lytic Polysaccharide Monooxygenases (LPMOs). *Biochem. Soc. Trans.* **2018**, *46* (6), 1431–1447.

(82) Quinlan, R. J.; Sweeney, M. D.; Lo Leggio, L.; Otten, H.; Poulsen, J.-C. N.; Johansen, K. S.; Krogh, K. B. R. M.; Jørgensen, C. I.; Tovborg, M.; Anthonsen, A.; et al. Insights into the Oxidative Degradation of Cellulose by a Copper Metalloenzyme That Exploits Biomass Components. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (37), 15079–15084.

(83) Bissaro, B.; Røhr, Å. K.; Müller, G.; Chylenski, P.; Skaugen, M.; Forsberg, Z.; Horn, S. J.; Vaaje-Kolstad, G.; Eijsink, V. G. H. Oxidative Cleavage of Polysaccharides by Monocopper Enzymes Depends on H2O2. *Nat. Chem. Biol.* **2017**, *13* (10), 1123–1128.

(84) Hangasky, J. A.; Iavarone, A. T.; Marletta, M. A. Reactivity of O2 versus H2O2 with Polysaccharide Monooxygenases. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (19), 4915–4920.

(85) Isaksen, T.; Westereng, B.; Aachmann, F. L.; Agger, J. W.; Kracher, D.; Kittl, R.; Ludwig, R.; Haltrich, D.; Eijsink, V. G. H.; Horn, S. J. A C4-Oxidizing Lytic Polysaccharide Monooxygenase Cleaving Both Cellulose and Cello-Oligosaccharides. *J. Biol. Chem.* **2014**, *289* (5), 2632–2642.

(86) Bacik, J.; Mekasha, S.; Forsberg, Z.; Kovalevsky, A. Y.; Vaaje-kolstad, G.; Eijsink, V. G. H.; Nix, J. C.; Coates, L.; Cuneo, M. J.; Unkefer, J.; et al. Neutron and Atomic Resolution X - Ray Structures of a Lytic Polysaccharide Monooxygenase Reveal Copper-Mediated Dioxygen Binding and Evidence for N - Terminal Deprotonation. **2017**, 8–11.

(87) Wierenga, R. K.; Kapetaniou, E. G.; Venkatesan, R. Triosephosphate Isomerase: A Highly Evolved Biocatalyst. *Cell. Mol. Life Sci.* **2010**, *67* (23), 3961–3982.

(88) Alahuhta, M.; Wierenga, R. K. Atomic Resolution Crystallography of a Complex of Triosephosphate Isomerase with a Reaction-Intermediate Analog: New Insight in the Proton Transfer Reaction Mechanism. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), NA-NA.

(89) Go, M. K.; Koudelka, A.; Amyes, T. L.; Richard, J. P. Role of Lys-12 in Catalysis by Triosephosphate Isomerase: A Two-Part Substrate Approach. *Biochemistry* **2010**, *49* (25), 5377–5389.

(90) Venkatesan, R.; Alahuhta, M.; Pihko, P. M.; Wierenga, R. K. High Resolution Crystal Structures of Triosephosphate Isomerase Complexed with Its Suicide Inhibitors: The Conformational Flexibility of the Catalytic Glutamate in Its Closed, Liganded Active Site. *Protein Sci.* **2011**, *20* (8), 1387–1397.

(91) Nilsson, K.; Ryde, U. Protonation Status of Metal-Bound Ligands Can Be Determined by Quantum Refinement. *J. Inorg. Biochem.* **2004**, *98* (9), 1539–1546.

(92) Polyansky, A. A.; Kuzmanic, A.; Hlevnjak, M.; Zagrovic, B. On the Contribution of Linear Correlations to Quasi-Harmonic Conformational Entropy in Proteins. *J. Chem. Theory Comput.* **2012**, *8* (10), 3820–3829.

(93) Capelli, S. C.; Bürgi, H.-B.; Dittrich, B.; Grabowsky, S.; Jayatilaka, D. Hirshfeld Atom Refinement. *IUCrJ* **2014**, *1* (5), 361–379.

Acknowledgments

This thesis is about research, but my PhD studies could not have gone so smoothly without the multitude of people that helped me in one way or another.

I have had an incredible supervising team. **Ulf**, my main supervisor, thank you for teaching me so much, for always being there and for being patient when I made stupid mistakes. Also, thank you for making sure that my PhD was on track all the time. **Esko**, you have been a great co-supervisor. I appreciate all the crystallography lessons, all the interesting ideas we discussed and all the trips and lunches we had together.

A lot of people have been part of this group and I have tried to work with and learn from each of them. Thank you all for sharing scientific ideas. **Majda**, thank you for being both a friend and colleague. I've had a lot of fun playing board games and hanging out with you. Also, thanks for keeping me up to date with any courses and events and all the bureaucratic stuff that I usually miss. **Erik**, thanks for being a friend and a mentor. Encouraging me to continue in academia means a lot to me. **Lili**, we have published a lot of cool scientific work together. Thanks for working with me so much. **Geng**, thanks for all the little presents you brought from China. **Martin, Magne**, thanks for being good office-mates and not complaining about my loud music through my headphones. **Justin**, thanks for bringing new crystallographic knowledge to the group. **Adrian**, I never thought that I would find in the group someone with so many things in common. Thanks for inviting me to the MolMod conferences and for so many beers and music discussions. **Francesco**, I wish you could read this. You were the first person I worked with and also my first friend here. Thank you for everything.

A big thanks to all the people from different departments who I worked with as part of the DecRec project. **Maria,** thank you for being a friend and for all the nice events you found in Lund. **Olof,** thanks for all the Thursdays at the Biochemistry beer club. **Johan, Rohit, Kristoffer, Sven, Alexander** and all the other students from the project, it has been nice working with you. My thanks of course also to the PIs who have guided us in this project: **Derek, Mikael, Ulf N, Hakon, Stina.**

Thanks to all the Teokem department for welcoming me and making me feel at home. **Alexei**, thank you for all the board games, all the weekday lunches, all the

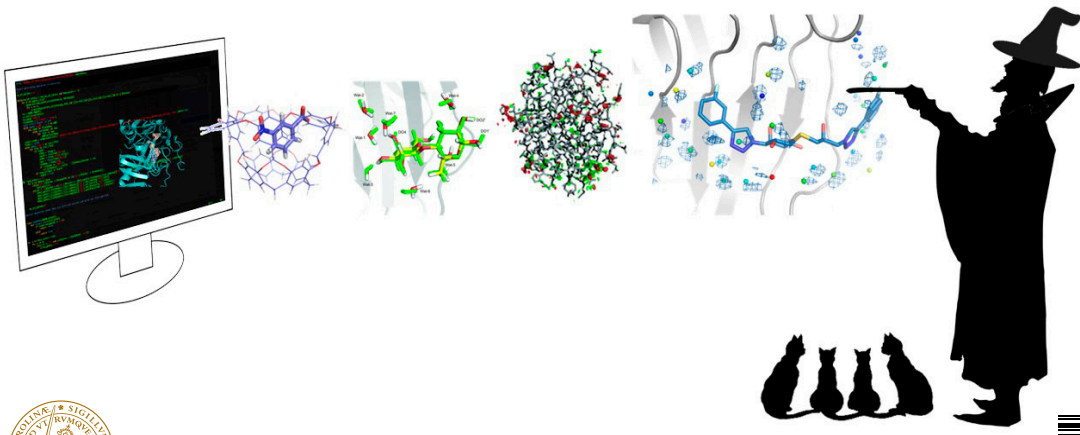
weekend lunches and in general all the fun we've had. My PhD would have been much more boring without you. **Valera, Per-Åke,** thanks for all the interesting discussions we had over lunch and for giving me the opportunity to teach. Thanks to the **Thursday board game group** for letting me annoy you with all my games for a while. Thanks to everybody who participated at fika for helping me pleasantly waste my time: **Stephanie, Ellen, Maria, Sam, Eric, Junhao, Dora, Joel, Amanda** and many more that I now forget.

Rujing, thank you for traveling to all the corners of the world with us. I hope we'll get to go to many more places in the future.

Multumiri si tuturor care nu au fost alaturi de mine in Lund dar care m-au sprijinit oricum. **Tata, Mama**, nu as fi putut ajuge pana aici fara voi. Multumesc ca mi-ati insuflat dragostea pentru stiinta si cercetare si pentru ca ati avut grija de mine pana nu am mai avut nevoie dar si mai tarziu. **Silvia,** multumesc pentru toate sfaturile despre cercetare si viata in general si pentru toate recomandarile de carti. Nu am reusit sa te ajung ca varsta dar macar am reusit sa ajung doctor ca tine. Sper ca o sa pot zice asta si cand o sa ajung profesor ca tine. Multumesc si celeilalte familii din care sunt parte acum: **Florin, Ana, Alexandru, Eugen, Roxana, Theo, Bobby si toti ceilalti** multumesc ca m-ati primit in familie.

Finally, **Ada**, thank you for feeding me, taking care of me and marrying me. You will get to be the last and foremost in all my acknowledgments from now on.

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903



LUND UNIVERSITY

ISBN: 978-91-7422-702-4

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University

9 789174 227024