# Quantum refinement in real and reciprocal space

Lundgren, Kristoffer

# Quantum refinement in real and reciprocal space

KRISTOFFER J. M. LUNDGREN | COMPUTATIONAL CHEMISTRY | LUND UNIVERSITY

```
def read_syst1(infile):
    qm_atoms = set()
    link_atoms = set()
    with open(infile, 'r') as file:
        line = file.readline()
        while line:
            atoms = parse_atoms_line(line)
            for atom
                link
            line = f
    return qm_atoms,

def write_pdb_h(outf
    hierarchy = mode
    atoms = hierarch
    for atom in atom
        atom_serial = int(atom.serial.strip())
        if atom.element_is_hydrogen() or atom_serial in link_pairs.keys():
            atom.element = ' H'
        if atom_serial in link_pairs.keys():
            c_qm = atoms[serial_to_index[link_pairs[atom_serial]]]
            atom.xyz = (c_qm.xyz[0] + g[atom_serial]*(atom.xyz[0] - c_qm.xyz[0]),
                c_qm.xyz[1] + g[atom_serial]*(atom.xyz[1] - c_qm.xyz[1]),
                c_qm.xyz[2] + g[atom_serial]*(atom.xyz[2] - c_qm.xyz[2]))
    hierarchy.write_pdb_file(file_name=outfile, crystal_symmetry=model.crystal_symmetry(), anisou=False)


def read_energy_and_gradient(infile):
    gradients = list()
    with open(infile, 'r') as file:
        line = file.readline()
        while line:
            if line == '# Number of atoms\n':
                file.readline()
                n_atoms = int(file.readline().strip())
            if line == '# The current total energy in Eh\n':
                file.readline()
                energy = float(file.readline().strip())
                break
            line = file.readline()
        for i in range(3): file.readline()
        for i in range(n_atoms):
            g = (float(file.readline()), float(file.readline()), float(file.readline()))
            gradients.append(g)
    return energy, gradients


def calculate_total_gradient(qm_gradients, mm1_gradients, mm_gradients, qm_atoms, g, link_pairs, serial_to_index):
    for atom in qm_atoms:
        mm_gradients[atom - 1] -= np.array(mm1_gradients[serial_to_index[atom]])
        if atom not in link_pairs.keys():
            mm_gradients[atom - 1] += np.array(qm_gradients[serial_to_index[atom]])
        else:
            mm_gradients[link_pairs[atom] - 1] += (1 - g[atom])*np.array(qm_gradients[serial_to_index[atom]])
            mm_gradients[atom - 1] += g[atom]*np.array(qm_gradients[serial_to_index[atom]])
    return mm_gradients


def rescale_qm_gradients(qm_gradients, w):
    return [tuple([w*component for component in gradient]) for gradient in qm_gradients]


def calculate_target(qm_energy, mm1_target, mm_target, w_qm):
    return mm_target - mm1_target + w_qm*qm_energy
```

Quantum refinement in real and reciprocal space

# Quantum refinement in real and reciprocal space

by Kristoffer J. M. Lundgren


LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Prof. Ulf Ryde, Dr. Esko Oksanen
Faculty opponent: Prof. Małgorzata Biczysko

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in lecture hall A, Kemicentrum, Lund, on Friday, the 12th of December 2025 at 13:00.

Organization
**LUND UNIVERSITY**

Department of Chemistry
Box 124
SE–221 00 LUND
Sweden

Document name
**DOCTORAL DISSERTATION**

Date of disputation
2025-12-12

Sponsoring organization

Author(s)
Kristoffer J. M. Lundgren

Title and subtitle
Quantum refinement in real and reciprocal space

Abstract
In order to understand and manipulate biological function at the molecular level, access to high-resolution structures of biological macromolecules is essential, as structure is intimately linked to function. The two main experimental techniques that can achieve atomic, or near-atomic, resolution are crystallography and electron microscopy. Unfortunately, the experimental data alone is typically not sufficient to obtain an accurate atomic model, owing to a poor data-to-parameter ratio. Therefore, prior knowledge of the chemical nature of the system is supplemented during refinement in the form of restraints. Traditional restraints are accurate for standard amino acids and nucleic acids, less so for novel ligands and metal sites. Additionally, transferability of these restraints is commonly assumed, ignoring the specific chemical environment. A solution is to use *in situ* quantum mechanical calculations for small, but interesting, parts of the structure, during refinement. Such an approach, called quantum refinement, has been shown to improve structures locally, to allow determination of protonation and oxidation states of ligands and metals and discriminate between different interpretations of the structure.

Previous implementations of quantum refinement have been limited to either X-ray or neutron crystallography data, using low-level quantum mechanical methods or not being able to treat metal sites. In this thesis, we present a new implementation of quantum refinement, called QRef. QRef supports X-ray, neutron and electron diffraction data, as well as cryo-EM data, and can use a wide range of quantum mechanical methods, allowing treatment of metal sites. QRef is released under a permissive license.

We have subsequently applied QRef to a variety of challenging biological systems. For a recent crystal structure of Fe-nitrogenase, we determined the protonation state of the homocitrate ligand. We performed a critical evaluation of metal sites in three cryo-EM structures of particulate methane monooxygenase, showing that several sites were incorrectly interpreted. This was the first time that quantum refinement had been applied to metal sites in cryo-EM structures and we suggest that quantum refinement is the method of choice for such systems. Furthermore, we have applied quantum refinement to data from XFEL and electron diffraction experiments, showing that quantum refinement can improve structures and interpretations in these cases as well. Finally, we have applied QRef to data from neutron diffraction experiments of manganese superoxide dismutase, demonstrating the challenges associated with neutron data of limited resolution.

Key words
quantum refinement, X-ray crystallography, neutron crystallography, electron crystallography, cryo-EM, nitrogenase, particulate methane monooxygenase, ribonucleotide reductase, manganese superoxide dismutase

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-8096-130-1 (print)
978-91-8096-131-8 (pdf)

Recipient's notes

Number of pages
222

Price

Security classification

# Quantum refinement in real and reciprocal space

by Kristoffer J. M. Lundgren


LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

**Cover illustration front:** The most central parts of the code in the `QRef` interface. Generated by using the plugin `CodeSnap` in `Visual Studio Code` together with the `Tokyo Night` theme, assembled and adjusted in `Microsoft PowerPoint`.

**Cover illustration back:** QR code with a link to the GitHub repository for `QRef`, generated using the `qrcode` and `PyPNG` modules in Python.

**Funding information:** This thesis work has been financially supported by grants from the Swedish Research Council (projects 2020-06176 and 2022-04978) and the Swedish Agency for Economic and Regional Growth ("Automated software pipeline for NMX data processing"). The computations were enabled by resources provided by LUNARC, The Centre for Scientific and Technical Computing at Lund University.



Faculty of Science, Department of Chemistry, Division of Computational Chemistry

ISBN: 978-91-8096-130-1 (print)
ISBN: 978-91-8096-131-8 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

*"Truth is much too complicated to allow anything but approximations."*
— John von Neumann

*"All models are wrong, but some are useful."*
— George E. P. Box

*Dedicated to the memory of my mother, Lena Lundgren (1955–2011), and my grandfather, Göte Karlsson (1928–2016). I miss you dearly.*

# Contents


List of publications . . . iii
Publications not included in this thesis . . . iv
Author contributions . . . v
Acknowledgements . . . vi
Abbreviations . . . viii
Popular science summary in English . . . x
Populärvetenskaplig sammanfattning på svenska . . . xv

**1 Introduction** **1**

**2 Structural biology** **3**
2.1 Crystallography . . . 4
2.1.1 Crystals . . . 4
2.1.2 Diffraction . . . 5
2.1.3 Atomic scattering factors . . . 7
2.1.4 Structure factors . . . 9
2.1.5 Atomic displacement and occupancy . . . 10
2.1.6 Bulk solvent . . . 11
2.1.7 The phase problem . . . 12
2.2 Cryogenic electron microscopy . . . 13
2.2.1 Image formation . . . 14
2.2.2 The contrast transfer function . . . 16
2.2.3 Image processing and 3D reconstruction . . . 17
2.3 Refinement . . . 18
2.3.1 The weight factor . . . 22

**3 Computational chemistry** **25**
3.1 Quantum chemistry . . . 25
3.1.1 The Schrödinger equation . . . 25
3.1.2 Hartree–Fock theory . . . 27
3.1.3 Basis sets . . . 30
3.1.4 Density functional theory . . . 32
3.2 Molecular mechanics . . . 35
3.3 Hybrid methods . . . 36

3.3.1 The link atom approach . . . 39
3.3.2 Choosing the model region . . . 40

**4 Quantum refinement 41**
4.1 Force fields in refinement . . . 41
4.2 Quantum refinement . . . 42
4.3 QRef . . . 44
4.3.1 Implementation details . . . 46
4.3.2 Manual restraints . . . 48
4.3.3 Workflow . . . 49
4.4 Validation metrics . . . 52

**5 Studied proteins 57**
5.1 Nitrogenase . . . 57
5.2 Manganese superoxide dismutase . . . 58
5.3 Particulate methane monooxygenase . . . 59
5.4 Ribonucleotide reductase . . . 62

**6 Summary of papers 65**
6.1 Paper I . . . 65
6.2 Paper II . . . 69
6.3 Paper III . . . 74
6.4 Paper IV . . . 76
6.5 Paper V . . . 78

**7 Conclusions and Outlook 81**

**References 85**

**Scientific publications 115**

# List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Protonation of Homocitrate and the $E_1$ State of Fe-Nitrogenase Studied by QM/MM Calculations**
H. Jiang, **K. J. M. Lundgren**, and U. Ryde
*Inorg. Chem.*, 62(48), 19433–19445, Nov. 2023. doi:10.1021/acs.inorgchem.3c02329.

II **Quantum refinement in real and reciprocal space using the Phenix and ORCA software**
**K. J. M. Lundgren**, O. Caldararu, E. Oksanen, and U. Ryde
*IUCrJ*, 11(6), 921–937, Nov. 2024. doi: 10.1107/S2052252524008406.

III **Critical evaluation of three cryo-EM structures of pMMO by quantum refinement**
G. Yuvaraj, **K. J. M. Lundgren**, E. Veenman, E. Oksanen, and U. Ryde
*Acta Cryst. D*, 81(11), 605–620, Nov. 2025. doi: 10.1107/S2059798325008356.

IV **Quantum refinement with electron diffraction and X-ray free-electron laser data: Comparative study of ribonucleotide reductase dimetal site**
**K. J. M. Lundgren**, X. Sun, L. Pacoste, R. Kumar, G. Hofer, H. Xu, X. Zou, M. Högbom, E. Oksanen, and U. Ryde
*J. Appl. Cryst.*, submitted.

V **Critical evaluation of two neutron structures of Mn superoxide dismutase with quantum refinement**
**K. J. M. Lundgren**, J. Bergmann, E. Oksanen, and U. Ryde
*J. Biol. Inorg. Chem.*, submitted.

All papers are published with open access.

## Publications not included in this thesis

VI **The reaction mechanism of iron and manganese superoxide dismutases studied by theoretical calculations**
L. Rulíšek, K. P. Jensen, **K. Lundgren**, and U. Ryde
*J. Comput. Chem.*, 27(12), 1398–1414, June 2006. doi: 10.1002/jcc.20450.

VII **The $Cu_B$ site in particulate methane monooxygenase may be used to produce hydrogen peroxide**
**K. J. M. Lundgren**, L. Cao, M. Torbjörnsson, E. D. Hedegård, and U. Ryde
*Dalton Trans.*, 54(8), 3141–3156, Feb. 2025. doi: 10.1039/D4DT03301A.

VIII **Can ferric-oxyl excited states explain elongated iron–oxygen bonds in heme peroxidase catalytic intermediates?**
L. J. Williams, J. J. A. G. Kamps, A. M. V. Branzanic, M. Lehene, **K. J. M. Lundgren**, K. Chatterjee, M. A. Doyle, P. S. Simon, H. Makita, A. J. Thompson, A. S. Brewster, T. Zhou, M. Lucic, M. T. Wilson, P. Aller, J. Sanchez Weatherby, L. Gee, S. Dehe, J. Yano, V. K. Yachandra, M. A. Hough, U. Ryde, A. M. Orville, J. F. Kern, R. L. Silaghi-Dumitrescu, and J. A. R. Worrall
*Nature Commun.*, submitted.

# Author contributions

I **Protonation of Homocitrate and the $E_1$ State of Fe-Nitrogenase Studied by QM/MM Calculations**

I developed and tested all code for the `QRef` interface. I performed and analysed all the quantum refinement calculations. I participated in the evaluation of the results. I participated in the writing and revision of the manuscript.

II **Quantum refinement in real and reciprocal space using the Phenix and ORCA software**

I developed and tested all code for the `QRef` interface. I performed and analysed all calculations. I participated in the evaluation of the results. I made all figures. I wrote the first draft of the methods section. I participated in the writing and revision of the manuscript.

III **Critical evaluation of three cryo-EM structures of pMMO by quantum refinement**

I extended the `QRef` interface to be compatible with real-space refinement. I supervised G. Y. and E. V., who performed most of the calculations and analysis. I performed some of the calculations. I participated in the evaluation of the results. I participated in the writing and revision of the manuscript.

IV **Quantum refinement with electron diffraction and X-ray free-electron laser data: Comparative study of ribonucleotide reductase dimetal site**

I developed the `QRef` interface that made it possible to extend quantum refinement to electron diffraction. I supervised X. S. I performed and analysed almost all calculations. I participated in the evaluation of the results. I made all figures. I participated in the writing of the manuscript.

V **Critical evaluation of two neutron structures of Mn superoxide dismutase with quantum refinement**

I extended the `QRef` interface to be able to handle symmetry interactions. I performed and analysed all calculations. I made all figures. I participated in the evaluation of the results. I participated in the writing of the manuscript.

# Acknowledgements

Here we are, four years later and my thesis is about to be submitted for printing. It has been a long journey, but I still remember that summer day in August 2021 when I returned to KC after sixteen years away. The only part left now is to write the Acknowledgements, which is a bit tricky. You want to thank everyone that has helped you along the way, but at the same time you do not want to forget anyone important. If I forget to mention you here, please forgive me, it is not intentional.

First and foremost, I want to thank my supervisor **Ulf** for accepting me as a PhD student and for all the support and guidance you have given me. Never before have I met anyone with such a fervour for science. I have also lost count on how many times you have cheered me up when I have been down, both scientifically and personally. Thank you. I could not have asked for a better supervisor. I would also like to apologise; I will probably never learn when to use "-" and when to use "–".

I would also like to thank **Esko** for being a great co-supervisor, always ready to help out with both scientific and practical matters (I still need help finishing moving). Thank you for all the lunches and beers over the years, they have been very much appreciated. Being introduced to the crystallographic community has been invaluable. I have also enjoyed our discussions, not only about science.

**Justin**, thank you for teaching me the basics of quantum refinement when I just started. And also for staying in the left lane when giving me a lift to Nottingham.

**Hao**, you were the perfect office mate. Two introverts, working in silence. Thank you **Joel** for the (forced) rubber duck debugging sessions, they helped me a lot. Thank you **Vihelm** for sending me funny videos to watch when I was sick and stuck at home. **Simon**, thank you for introducing me to pastis. Thank you **Iria** for not stealing my stuff (as far as I have noticed). **Xiaoli**, I hope we meet again and thank you for the amazing dumplings. Thank you **Ernst** for all the discussions over fika. Never shave your beard or you will lose all your powers. Thank you **Mickael** for being a good friend. I hope Links/Lynx treats you well. **Isabel**, thank you for organising so many social events and always being a supportive friend. Likewise, **David**, thank you for being a great friend and always having time to talk about pretty much everything. **Marcos**, thank you for all the humour and allowing me to bounce ideas off of you. Thank you **Valera** and **Magnus** for the technical support and all the nice chats we have had. Thank you **Nikol**, you were the best co-host at the Christmas party. **Ismail**, thank you for the support during the final stretch. I wish you both the best of luck with your upcoming defenses. **Gayathri**, thank you for helping me out with graphics and illustrations, and continuing on the gospel of quantum refinement.

**Magne**, **Victor**, **Georgios**, **Kosala**, **Gaia**, **Andreas**, **Xiaofan**, **Diletta** and all other current and former members of the Journal Club, thank you for sharing and discussing interesting papers. **Jens**, thank you for the coffees, they have been a lifesaver on several occasions. Thank you **Mikael** for the programming discussions. I will learn Rust at some point, I promise. **Derek**, thank you for teaching me the basics of structural biology. **Elija**, thank you for stress testing my code.

**Kalle**, my dear friend, I do not even know where to start. You are always there for me, no matter what. I wish you and your family all the best in the future. Likewise, **Oskar**, I know that I can always count on you.

Almost last, but definitely not least (quite the opposite, actually), **Susanne**, thank you for all your support and for sharing your life with me. You are and have been my safe haven throughout not only this journey, but also through all other ups and downs. I love you and I look forward to spending the rest of my life with you.

Finally, I want to give a shout-out to the excellent command line tool `grep`, without which this thesis would not have been possible. Thank you, `grep`.

# Abbreviations

| | |
|---|---|
| **2D** | Two-Dimensional |
| **3D** | Three-Dimensional |
| **ADP** | Atomic Displacement Parameter |
| **AMBER** | Assisted Model Building with Energy Refinement |
| **API** | Application Programming Interface |
| **ASE** | Atomic Simulation Environment |
| **BS** | Broken-Symmetry |
| **cctbx** | Computational Crystallography ToolboX |
| **CDL** | Conformation Dependent Library |
| **CNS** | Crystallography and NMR System |
| **CPCM** | Conductor-like Polarizable Continuum Model |
| **cryo-EM** | cryogenic Electron Microscopy |
| **CTF** | Contrast Transfer Function |
| **DFT** | Density Functional Theory |
| **DNA** | DeoxyriboNucleic Acid |
| **ED** | Electron Diffraction |
| **EM** | Electron Microscopy |
| **EMDB** | Electron Microscopy Data Bank |
| **EPR** | Electron Paramagnetic Resonance |
| **EXAFS** | Extended X-ray Absorption Fine Structure |
| **GGA** | Generalized Gradient Approximation |
| **GTO** | Gaussian-Type Orbital |
| **HAR** | Hirshfeld Atom Refinement |
| **HF** | Hartree–Fock |
| **IAM** | Independent Atom Model |
| **IUCr** | International Union of Crystallography |
| **KS** | Kohn–Sham |
| **L-BFGS** | Limited-memory Broyden–Fletcher–Goldfarb–Shanno |
| **LCAO** | Linear Combination of Atomic Orbitals |
| **LDA** | Local Density Approximation |
| **MAD** | Multi-wavelength Anomalous Diffraction |
| **MD** | Molecular Dynamics |
| **MicroED** | Microcrystal Electron Diffraction |
| **MIR** | Multiple Isomorphous Replacement |
| **MM** | Molecular Mechanics |
| **MnSOD** | Manganese SuperOxide Dismutase |
| **MR** | Molecular Replacement |

| | |
|---|---|
| **ND** | Neutron Diffraction |
| **NMR** | Nuclear Magnetic Resonance |
| **ONIOM** | Our own N-layered Integrated molecular Orbital and Molecular mechanics |
| **PDB** | Protein Data Bank |
| **PES** | Potential Energy Surface |
| **PHENIX** | Python-based Hierarchical ENvironment for Integrated Xtallography |
| **PME** | Particle Mesh Ewald |
| **pMMO** | particulate Methane MonoOxygenase |
| **PSF** | Point Spread Function |
| **QM** | Quantum Mechanics |
| **QM/MM** | Quantum Mechanics/Molecular Mechanics |
| **QR** | Quantum Refinement |
| **RNR** | RiboNucleotide Reductase |
| **RSCC** | Real-Space Correlation Coefficient |
| **RSR** | Real-Space R |
| **RSZD** | Real Space Z-Difference |
| **SAD** | Single-wavelength Anomalous Diffraction |
| **SCF** | Self-Consistent Field |
| **SD** | Slater Determinant |
| **SEM** | Scanning Electron Microscopy |
| **sMMO** | soluble Methane MonoOxygenase |
| **SNR** | Signal-to-Noise Ratio |
| **SOD** | Super-Oxide Dismutase |
| **SPA** | Single Particle Analysis |
| **STO** | Slater-Type Orbital |
| **SQM** | Semiempirical Quantum Mechanics |
| **TEM** | Transmission Electron Microscopy |
| **XFEL** | X-ray Free Electron Laser |
| **XRD** | X-Ray Diffraction |

## Popular science summary in English

All life as we know it depends on something called "proteins" in one way or another. We also know that life depends on water, DNA, RNA, lipids, carbohydrates and a number of other molecules, but proteins seem to be the most versatile and important ones.

Proteins are large molecules, consisting of smaller building blocks called "amino acids", which in turn are made up of atoms. These amino acids are connected to each other in long chains, that fold into complex shapes. The specific sequence of amino acids in a protein determines its unique three-dimensional structure and function.

Proteins play a crucial role in virtually all biological processes, including catalysing metabolic reactions, replicating DNA and transporting molecules. Catalysis is the process of speeding up chemical reactions without being consumed in the process and proteins that do this are called "enzymes". Enzymes are essential for life, as they allow chemical reactions to occur at the kind of temperatures and pressures that you and I experience every day. Nature has spent billions of years perfecting the art of using proteins as catalysts through evolution. Some of the reactions that enzymes catalyse are reactions that we humans have started to perform on an industrial scale, by using toxic chemicals or at extreme temperatures and pressures, spending a lot of energy. This has been the only way we have managed to perform these reactions. Imagine if we could just mimic what nature is already capable of doing in our industrial processes!

Understanding how proteins work is also beneficial from a medical perspective. Many diseases are caused by proteins that do not function properly. In other cases, proteins may become misfolded or aggregated, leading to conditions such as Alzheimer's disease or cystic fibrosis. Or perhaps we want to shut down specific proteins that are essential for the survival of harmful bacteria or viruses. If we understand how the involved proteins work, we can design drugs that specifically target them and inhibit their function.

And on top of all that, humans are curious creatures, always wanting to understand how things work. So it is no surprise that proteins are of great interest to scientists in many different fields.

Interestingly, research has shown that structure is usually more important than the sequence of amino acids for a protein's function. This means that even if two proteins have very similar sequences, they can have different functions if their structures are different. Conversely, proteins with different sequences can have similar functions if their structures are alike. It seems then that we need to know the structure of proteins in order to understand how they work.

A multitude of experimental techniques exist to study protein structures, but if we want a resolution that allows us to see the individual atoms, we typically need to use techniques called "crystallography" or "cryo-electron microscopy".

In crystallography, scientists grow crystals of the protein they want to solve the structure of. A crystal is a material where the atoms are arranged in a highly ordered and periodically repeating pattern. The scientists then expose the crystal to a beam of radiation (X-rays, neutrons or electrons). The ordering in the crystal causes the incoming radiation to "diffract" in specific directions, creating a pattern of spots on the detector, as well as amplifying the signal. This pattern is a fingerprint of how the atoms are arranged inside the crystal. Using words from the language of mathematics, the pattern is a "Fourier transform" of some kind of density inside the crystal. Using more fancy words from mathematics, we can also say that we look at the data in "reciprocal space" instead of "real space". With a bit more math (and computers) we can then turn that fingerprint into a three-dimensional density map. Depending on the type of radiation used, the density is either an electron density (X-rays), a nuclear density (neutrons) or an electrostatic potential (electrons). This density is what we ultimately want to know from the experiment.

You might not have noticed it, but there is a big if in the preceding paragraph; growing crystals of proteins is not a trivial task. In fact, it is often the biggest bottleneck in crystallography. Soluble proteins (those that like to hang out in water) are typically easier to crystallise than membrane proteins (those that like to hang out in oily environments, such as cell membranes). On top of that, a protein crystal is not necessarily representative of the protein in its natural environment, as the crystal packing might distort the protein structure somewhat. What if we could just use a microscope with a very powerful lens to look at the protein in its native state directly, without the need to grow crystals?

Using a very high voltage, we can accelerate electrons to very high speeds, giving them a very short wavelength, which enables very high resolution imaging. This is exploited in cryogenic electron microscopy (cryo-EM), where a beam of electrons is used to image the protein directly. The scientists can then take many pictures of the protein, embedded in a thin layer of vitrified ice to preserve its native state, in different orientations. Because actual images are taken, cryo-EM data is said to be in "real space". This is not necessarily entirely true, as the (electromagnetic) lenses used may or may not introduce Fourier transforms back and forth (and additionally, Fourier transforms are sometimes used in the image processing steps), but nonetheless, the data is fundamentally different from crystallographic data and considered to be in real space. The scientists then use computers (and more math) to combine these pictures into a three-dimensional density, which again is what we want to know from the experiment. In cryo-EM, this density is an electrostatic potential, similar to what we get from electron diffraction experiments.

Unfortunately, the images from cryo-EM are very noisy, particularly for biological samples. This is because proteins are very sensitive to radiation damage from the electrons and this limits the resolution. While crystallography routinely reaches atomic level resolution, cryo-EM is not quite there yet, although the technique is rapidly improving.

With a density in hand, be it from crystallography or cryo-EM, the next step is to build a model of the protein that best explains the density. When the model is next to complete, an automated process called "refinement" is used to improve the model as much as possible by varying the parameters of the model. Refinement can thus be seen as an optimisation process. Unfortunately, the data from the experiments is often not of sufficient quality to be used on its own during refinement of the model, because the data-to-parameter ratio is too low. If one refines a model and uses only the experimental data, the resulting model is usually overfitted to the data and contains many unrealistic features such as too short or too long bond lengths, distorted angles, atoms being too close to each other, etc. To alleviate this problem, additional information is added during refinement in the form of what is called "restraints". These restraints are typically based on empirical observations or calculated properties of similar systems and then assumed to be transferable to the system at hand. From a mathematical perspective, these restraints can be seen as additional terms in the optimisation process that penalise unrealistic features in the model. From a probabilistic perspective, the restraints are prior information that is combined with the experimental data (the likelihood) to give a posterior probability distribution of the model parameters.

The restraints can in principle be anything, but are usually in the form of geometric restraints (bond lengths, angles, torsions, planarity, chirality, etc.) or chemical information (e.g. known binding motifs, secondary structure elements, etc.). But by categorising the restraints like this, we overlook much of the underlying physics that govern how atoms and molecules interact with each other. Especially when proteins (in particular enzymes) tend to cheat a little bit and use not only standard amino acids, but also metal ions or unusual cofactors (small molecules that help the protein) to perform their function, which may not be well described by simple geometric restraints.

What if we could use the laws of physics directly to describe how the atoms in the protein interact with each other for the restraints? This is where quantum mechanics comes in.

Quantum mechanics is the most fundamental theory we have to describe how atoms and molecules behave and relies on a set of axioms (axioms are things that we just accept to be true, irregardless if they are or not) that have been experimentally verified to a high degree of accuracy. So for now, it is probably safe to say that quantum mechanics is the best de-

scription of nature that we have.

In principle, quantum mechanics can describe everything (that we can observe) about any physical system, which obviously includes proteins. In practise, a *very* complicated equation has to be solved, called the Schrödinger equation, in order to get the quantum mechanical description of a system. This equation can only be solved exactly for the simplest systems, such as the hydrogen atom. For more complicated systems, we have to resort to some approximations in order to make the calculations feasible. Nevertheless, quantum mechanics still provides a more accurate description of the system, even when using these approximations, than assuming that a certain bond exists with an ideal length or that an angle should be this or that.

However, quantum mechanical calculations suffer the problem of poor scaling with system size, which means that as the system gets larger, the calculations become rapidly more expensive. A way to work around this problem is to, for large systems, use quantum mechanics only for the most important part of the system (e.g. metals in an active site of an enzyme) and treat the rest of the system with a cheaper method (e.g. classical mechanics). In the language of computational chemistry, this is usually called "QM/MM" (Quantum Mechanics/Molecular Mechanics). I like to think of this more as a hybrid scheme between different levels of theory for different parts of the system. Though, the general idea is the same as QM/MM.

An approach like this, where quantum mechanical calculations are used during refinement of protein structures, is called "quantum refinement". An additional feature of quantum refinement, as to not just improve the parameters for a given model, is that quantum refinement can be used to test different models ("hypotheses") against each other, even when the experimental data alone is not sufficient to distinguish between them. This is possible because quantum mechanics implicitly includes all physical effects. For example, if we want to know the protonation state of a certain amino acid in the protein (i.e. if it has an extra hydrogen atom attached or not), this may not be directly visible in the experimental data. However, the protonation state will affect how the amino acid interacts with its surroundings, which in turn will affect the geometry and energy of the system. By testing different protonation states and comparing the resulting quantum-refined models, we can deduce which protonation state is the most likely one, even though the data alone is not sufficient to make this distinction. In a similar manner, quantum refinement can also be used to test other kinds of hypotheses, such as the oxidation state of a metal ion in an active site or the presence or absence of a water molecule, etc.

Quantum refinement has been around since 2002 and nowadays there exist several implementations of the method. In general, most existing implementations rely on an older refinement software, which is no longer being actively developed or maintained, or limits

the available quantum mechanical methods to lower accuracy methods, as well as often excluding treatment of metals completely. This limits the usability of quantum refinement, as modern macromolecular refinement software have many new and important features. Not being able to model the prior information to a sufficient degree of accuracy may lead to suboptimal results from quantum refinement.

As the central work of this thesis, I have implemented a new quantum refinement interface using modern macromolecular refinement software, which can be used together with arbitrary levels of quantum mechanical methods, for (almost) all kinds of elements. I named this interface `QRef`. Additionally, I have extended the kind of data sources to which quantum refinement can be applied, so that hybrid scheme quantum refinement now also applies to real-space data. `QRef` has been tested and thoroughly validated (**Paper II**). I have used `QRef` to identify the protonation state of a metal cofactor in a very complicated metalloenzyme called nitrogenase (**Paper I**). Nitrogenase is the enzyme responsible for nitrogen fixation, where inert nitrogen gas from the atmosphere is converted into bioavailable ammonia. Additionally, I have used `QRef` to apply quantum refinement to kinds of datasets where it has not been used before, e.g. to a membrane metalloprotein called particulate methane monooxygenase (pMMO), where the data originates from cryo-EM (**Paper III**). pMMO is an enzyme that converts methane into methanol. The location of the active site in pMMO still under debate. Through `QRef` I have also used quantum refinement for the first time with data from electron diffraction and X-ray free-electron lasers (**Paper IV**), for an enzyme that catalyses the formation of nucleotides for DNA. Finally, I have used `QRef` to evaluate two neutron diffraction structures of an important antioxidant enzyme, called manganese superoxide dismutase (**Paper V**).

## Populärvetenskaplig sammanfattning på svenska

Allt liv som vi känner det är på ett eller annat sätt beroende av något som kallas ”proteiner”. Vi vet också att liv är beroende av vatten, DNA, RNA, lipider, kolhydrater och ett antal andra molekyler, men proteiner verkar vara de mest mångsidiga och viktiga.

Proteiner är stora molekyler, bestående av mindre byggstenar som kallas ”aminosyror”, vilka i sin tur består av atomer. Dessa aminosyror är sammanlänkade i långa kedjor, som veckas till komplexa former. Den specifika sekvensen av aminosyror i ett protein bestämmer dess unika tredimensionella struktur och funktion.

Proteiner spelar en avgörande roll i praktiskt taget alla biologiska processer. T.ex. katalyserar de metaboliska reaktioner, replikerar DNA och transporterar molekyler. Katalys innebär processen att påskynda kemiska reaktioner utan att själv förbrukas och proteiner som gör detta kallas ”enzymer”. Enzymer är viktiga för allt liv, eftersom de möjliggör kemiska reaktioner vid rumstemperatur och normalt tryck, den typ av temperaturer och tryck som du och jag upplever varje dag. Naturen har ägnat miljarder år åt att fullända konsten att använda proteiner som katalysatorer genom evolutionen. Några av de reaktioner som enzymer katalyserar är reaktioner som vi människor har börjat utföra i industriell skala, men ofta med hjälp av giftiga kemikalier eller vid mycket höga temperaturer och tryck, vilket förbrukar mycket energi. Detta är det enda sätt som vi hittills har lyckats utföra dessa reaktioner i industriell skala. Tänk om vi i stället kunde härma hur naturen utför dessa processer!

Att förstå hur proteiner fungerar är också fördelaktigt ur ett medicinskt perspektiv. Många sjukdomar orsakas av proteiner som inte fungerar korrekt eller som har blivit felveckade eller aggregerade, vilket leder till sjukdomar som Alzheimers och cystisk fibros. En annan möjlighet är att vi vill stänga av specifika proteiner som är avgörande för överlevnaden av skadliga bakterier eller virus. Om vi förstår hur de involverade proteinerna fungerar kan vi designa läkemedel som specifikt riktar sig mot dem och hämmar deras funktion.

Dessutom är människor nyfikna varelser som alltid vill förstå hur saker fungerar, så det är ingen överraskning att proteiner är av stort intresse för forskare inom många olika områden.

Intressant nog har forskning visat att strukturen vanligtvis är viktigare än sekvensen av aminosyror för ett proteins funktion. Det betyder att även om två proteiner har mycket liknande sekvenser så kan de ha olika funktioner om deras strukturer skiljer sig. Omvänt så kan proteiner med olika sekvenser ha liknande funktioner om deras strukturer är lika. Därför behöver vi känna till proteinernas struktur för att förstå hur de fungerar.

Det finns en mängd experimentella tekniker för att studera proteinstrukturer, men om vi vill ha en upplösning som gör att vi kan se de enskilda atomerna behöver vi vanligtvis använda de tekniker som kallas för ”kristallografi” eller ”kryoelektronmikroskopi”.

Inom kristallografi odlar forskare kristaller av det protein de är intresserade av. En kristall är ett material där atomerna är arrangerade i ett välordnat och periodiskt upprepat mönster. Forskarna exponerar sedan kristallen för strålning (röntgenstrålar, neutroner eller elektroner). Ordningen i kristallen förstärker signalen och får den inkommande strålningen att ”diffraktera” i specifika riktningar, vilket skapar ett mönster av fläckar på detektorn. Detta mönster är ett fingeravtryck för hur atomerna är arrangerade inuti kristallen. På matematikens språk säger vi att mönstret är en ”Fouriertransform” av någon sorts täthet inuti kristallen. Med fler fina ord från matematiken kan vi också säga att vi tittar på den experimentella datan i det ”reciproka rummet” istället för det ”verkliga rummet”. Med lite mer matematik (och datorer) kan vi sedan omvandla fingeravtrycket till en tredimensionell täthetskarta. Beroende på vilken typ av strålning som används så är tätheten antingen en elektrontäthet (röntgenstrålar), en kärntäthet (neutroner) eller en elektrostatisk potential (elektroner). Denna täthet är vad vi i slutändan vill veta från experimentet.

Du har kanske inte märkt det, men det finns ett stort om i föregående stycke; att odla proteinkristaller är ingen enkel uppgift. Faktum är att det ofta är den största flaskhalsen inom kristallografi. Lösliga proteiner (de som gillar att hänga i vatten) är vanligtvis lättare att kristallisera än membranproteiner (de som gillar att hänga i feta miljöer, såsom cellmembran). Dessutom är en proteinkristall inte nödvändigtvis representativ för proteinet i dess naturliga miljö, eftersom kristallpackningen kan förvränga proteinstrukturen något. Tänk om vi i stället kunde använda ett mikroskop med en mycket kraftfull lins för att titta direkt på proteinet i dess ursprungliga tillstånd, utan att behöva odla kristaller?

Med en mycket hög spänning kan vi accelerera elektroner till mycket höga hastigheter, vilket ger dem en mycket kort våglängd och möjliggör avbildning med mycket hög upplösning. Detta utnyttjas i kryoelektronmikroskopi (kryo-EM), där en elektronstråle används för att avbilda proteinet direkt. Forskarna kan sedan ta många bilder av proteinet, inbäddade i ett tunt lager av vitrifierad is för att bevara dess ursprungliga tillstånd, i olika orienteringar. Eftersom faktiska bilder tas, sägs kryo-EM-data vara i ”verkliga rymden”. Detta är inte nödvändigtvis helt sant, eftersom de (elektromagnetiska) linserna som används kan introducera Fouriertransformationer fram och tillbaka (Fouriertransformationer används dessutom ibland i bildbehandlingssteget), men datan skiljer sig fundamentalt från kristallografisk data och anses vara i verkliga rymden. Forskarna använder sedan datorer (och mer matematik) för att kombinera dessa bilder till en tredimensionell täthet, vilket återigen är vad vi vill få fram från experimentet. I kryo-EM är denna täthet en elektrostatisk potential, liknande den vi får från elektrondiffraktionsexperiment.

Tyvärr är bilderna från kryo-EM mycket brusiga, särskilt för biologiska prover. Detta beror på att proteiner är mycket känsliga för strålningsskador från elektronerna och detta begränsar upplösningen. Medan kristallografi rutinmässigt når atomär upplösning, är kryo-EM inte riktigt där än, även om tekniken snabbt förbättras.

Med en täthetskarta, vare sig den kommer från kristallografi eller kryo-EM, så är nästa steg att bygga en modell av proteinet som bäst förklarar kartan. När modellen är nästan färdigbyggd används en automatiserad process som kallas ”förfining” för att förbättra modellen så mycket som möjligt genom att variera modellens parametrar. Förfining kan således ses som en optimeringsprocess. Tyvärr är datan från experimenten ofta inte av tillräckligt hög kvalitet för att användas på egen hand vid förfiningen av modellen, eftersom data/parameterförhållandet är för lågt. Om man förfinar en modell och endast använder experimentell data så blir den resulterande modellen vanligtvis överanpassad mot datan och uppvisar många orealistiska särdrag, t.ex. för korta eller för långa bindningslängder, förvrängda vinklar, atomer som är för nära varandra, etc. För att lindra detta lägger man till ytterligare information under förfiningen i form av vad som kallas ”begränsningar”. Dessa begränsningar är vanligtvis baserade på empiriska observationer eller beräknade egenskaper hos liknande system som sedan antas vara överförbara till det aktuella systemet. Ur ett matematiskt perspektiv kan dessa begränsningar ses som ytterligare termer i optimeringsprocessen som bestraffar orealistiska egenskaper i modellen. Ur ett sannolikhetsperspektiv är begränsningarna förhandsinformation som kombineras med experimentell data (den betingade sannolikheten) för att ge en posterior sannolikhetsfördelning av modellparametrarna.

Begränsningarna kan i princip vara vad som helst, men är vanligtvis i form av geometriska begränsningar (bindningslängder, vinklar, torsioner, planaritet, kiralitet, etc.) eller kemisk information (t.ex. kända bindningsmotiv, sekundära strukturelement, etc.). Men om man kategoriserar begränsningarna på detta sätt missar vi mycket av den underliggande fysiken som styr hur atomer och molekyler interagerar med varandra, speciellt när proteiner (i synnerhet enzymer) tenderar att fuska lite och inte bara använder sig av vanliga aminosyror, utan också metalljoner eller ovanliga kofaktorer (små molekyler som hjälper proteinet) att utföra sin funktion, vilka ofta inte beskrivs väl av enkla geometriska begränsningar.

Tänk om vi i stället kunde använda fysikens lagar direkt för att beskriva hur atomerna i proteinet interagerar med varandra för begränsningarna? Det är här kvantmekaniken kommer in i bilden.

Kvantmekanik är den mest grundläggande teorin vi har för att beskriva hur atomer och molekyler beter sig, och den bygger på en uppsättning axiom (axiom är saker som vi helt enkelt accepterar som sanna, oavsett om de är det eller inte) som har verifierats experimentellt med mycket hög noggrannhet. För närvarande är kvantmekanik den bästa beskrivningen av naturen som vi har.

I princip kan kvantmekanik beskriva allt (som vi kan observera) om vilket fysiskt system som helst, vilket uppenbarligen inkluderar proteiner. I praktiken måste en *mycket* komplicerad ekvation lösas, Schrödinger-ekvationen, för att erhålla en kvantmekanisk beskrivning av ett system. Denna ekvation kan bara lösas exakt för de allra enklaste systemen, såsom väteatomen. För mer komplicerade system måste vi använda en del approximationer för att göra beräkningarna genomförbara. Ändå ger kvantmekaniken fortfarande en bättre beskrivning av systemet, även när man använder dessa approximationer, än att anta att en viss bindning existerar med en ideal längd eller att en vinkel ska vara si eller så.

Tyvärr skalar kvantmekaniska beräkningar mycket dåligt med avseende på storleken på det studerade systemet, vilket innebär att när systemet blir större tar beräkningarna snabbt mycket lång tid att utföra. Ett sätt att komma runt detta problem för stora system är att använda kvantmekanik endast för den viktigaste delen av systemet (t.ex. metaller i ett enzyms reaktiva centrum) och behandla resten av systemet med en billigare metod (t.ex. klassisk mekanik). I beräkningskemins språk kallas detta vanligtvis ”QM/MM” (kvantmekanik/molekylmekanik). Jag föredrar att tänka på detta mer som ett hybridschema mellan olika teorinivåer för olika delar av systemet. Den generella idén är dock densamma som för QM/MM.

En metod som denna, där kvantmekaniska beräkningar används under förfiningen av proteinstrukturer, kallas för ”kvantförfining”. En ytterligare egenskap hos kvantförfining, alltså inte bara att förbättra parametrarna för en given modell, är att kvantförfining kan användas för att jämföra olika modeller (”hypoteser”) mot varandra, även när den experimentella datan inte ensam kan skilja dem åt. Detta är möjligt eftersom kvantmekanik implicit inkluderar alla fysiska effekter. Om vi t.ex. vill urskilja protoneringstillståndet för en viss aminosyra i ett protein (dvs. om den har en extra väteatom eller inte), kanske detta inte är direkt synligt i den experimentella datan. Protoneringstillståndet kommer dock att påverka hur aminosyran interagerar med sin omgivning, vilket i sin tur kommer att påverka systemets geometri och energi. Genom att testa olika protoneringstillstånd och jämföra de resulterande kvantförfinade modellerna kan vi avgöra vilket protoneringstillstånd som är mest sannolikt, även om datan ensam inte är tillräcklig för att avgöra detta. På liknande sätt kan kvantförfining också användas för att testa andra typer av hypoteser, t.ex. oxidationstillståndet för en metalljon i det reaktiva centrumet eller närvaron eller frånvaron av en vattenmolekyl.

Kvantförfining har funnits sedan 2002 och det finns flertalet implementeringar av metoden. I allmänhet förlitar sig de flesta befintliga implementeringar på en äldre förfiningsprogramvara, som inte längre aktivt utvecklas eller underhålls, eller använder kvantmekaniska metoder med låg noggrannhet, eller kan inte behandla metaller. Detta begränsar användbarheten av kvantförfining, eftersom modern förfiningsprogramvara har många nya och

viktiga funktioner. Att inte kunna modellera den tidigare informationen med tillräckligt hög noggrannhet kan leda till suboptimala resultat från kvantförfining.

Det centrala arbetet i denna avhandling har varit att implementera ett nytt kvantförfiningsgränssnitt som använder modern förfiningsprogramvara, som kan användas tillsammans med kvantmekanisk metodik på godtycklig nivå, för (nästan) alla typer av grundämnen. Jag gav detta gränssnitt namnet `QRef`. Dessutom har jag utökat den typ av datakällor där kvantförfining kan tillämpas, så att hybridschemakvantförfining nu även inkluderar data från kryo-EM. `QRef` har testats och noggrant validerats (**Artikel II**). Jag har använt `QRef` för att identifiera protoneringstillståndet för en metallkofaktor i ett mycket komplicerat metallenzym som kallas för nitrogenas (**Artikel I**). Nitrogenas är det enzym som ansvarar för kvävefixering genom omvandling av kvävgas från atmosfären till ammoniak, vilket gör kväve biotillgängligt. Dessutom har jag använt `QRef` för att utföra kvantförfining med nya typer av data, t.ex. för ett membranmetallprotein som kallas partikulärt metanmonooxygenas (pMMO), där datan kommer från kryo-EM (**Artikel III**). pMMO är ett enzym som omvandlar metan till metanol. Det är fortfarande oklart vilket som är det reaktiva centrumet i pMMO. Med `QRef` har jag också använt kvantförfining för första gången med data från elektrondiffraktion och röntgenfrielektronlaser (**Artikel IV**), för ett enzym som katalyserar bildandet av nukleotider för DNA. Slutligen har jag använt `QRef` för att granska två neutrondiffraktionsstrukturer hos ett viktigt antioxidantenzym som kallas mangansuperoxiddismutas (**Artikel V**).

# Chapter 1

# Introduction

Proteins are arguably among the most interesting molecules in nature as they are the building blocks of life and play essential roles in virtually all known biological processes. Apart from scientific curiosity regarding how proteins work, having a detailed understanding of their function is crucial for fields such as medicine, chemistry, biology and biotechnology. While proteins can be studied experimentally at various levels of detail, a comprehensive understanding of their structure at an atomic level is essential for elucidating their function. In turn this also allows for the development of rational tools to manipulate them, as well as providing starting models for computational exploration of the proteins.

Several techniques exist to gather information about protein structures, for example (ordered in roughly increasing resolution) proteolysis, mass spectroscopy, small–angle scattering, nuclear magnetic resonance spectroscopy, cryo-electron microscopy as well as diffraction experiments (employing either X-rays, neutrons or electrons). The latter two, cryo-electron microscopy and diffraction experiments allows for, or close to, atomic resolution. However, unless the data from experiments are of very high resolution, the data alone is not sufficient to produce accurate atomic models of the proteins due to a low data-to-parameter ratio. This often results in inconclusive or to a degree uncertain models of the protein under study. In order to alleviate this problem, prior knowledge about the system in question can be supplemented during the model building process. This prior information can in principle be anything, but is typically in the form of geometric restraints (bond lengths, angles, torsions, planarity, chirality, etc.) or chemical information (e.g. known binding motifs, secondary structure elements, etc.). Common to them all is that the parametrisation is based on empirical observations or calculations for similar, but not necessarily identical systems, and thus may not fully capture the underlying physics of the particular system of interest.

This thesis pertains to a method called "quantum refinement", which aims to improve the accuracy of protein models by incorporating *in situ* quantum mechanical calculations during the refinement process of the protein model, where the data is originates from either real space (cryo-electron microscopy) or reciprocal space (diffraction experiments). In theory, while accepting the axioms of quantum mechanics, quantum mechanics does produce a correct description of any physical system. In practice, as to not make the calculations unfeasible, approximations must be made, reducing the accuracy of the calculations somewhat but nonetheless yielding physically improved models of the proteins. Another aspect of quantum refinement is that the method also allows for hypothesis testing through implicit effects, making it possible to deduce what is the most likely structure of the protein being studied, even though the data alone is not sufficient to make these predictions.

The aim of this thesis is threefold: First, to implement quantum refinement using modern macromolecular refinement and quantum chemical software, as well as to establish protocols for performing quantum refinement and validate the performance. Second, to extend the use of quantum refinement to data sources where quantum refinement has not been applied before, again validating the approach. Third, to apply quantum refinement to interesting new systems.

In the following chapters, I will first present some of the underlying aspects of structural biology (chapter 2), followed by an overview of some of the tools used in computational chemistry (chapter 3). These two chapters will in a sense be merged in chapter 4, where I will describe the method of quantum refinement, as well as the implementation of quantum refinement that resulted from the work carried out during this thesis. In chapter 5 I will give a brief overview of the proteins that I have studied. Chapter 6 provides a summary of my publications. Lastly, in chapter 7, I will discuss conclusions and some possible extensions of quantum refinement.

# Chapter 2

# Structural biology

While structural biology is a quite broad field that encompasses many different aspects of biology and biochemistry, the field is primarily concerned with the three-dimensional structure determination of biological macromolecules. In turn, this also allows probing their functions and interactions (Liljas et al., 2016). The dominating technique used to obtain structural information for macromolecules is X-ray crystallography (XRD) with, as of the writing of this thesis, 197313 published structures in the Protein Data Bank (PDB), constituting 82 % of all entries (Berman et al., 2003, 2007; wwPDB Consortium, 2019, 2025b). Owing in part to the so called "resolution revolution", the number of cryo-electron microscopy (cryo-EM) structures has also increased dramatically over the last decade, with the Electron Microscopy Data Bank (EMDB) currently hosting 48984 entries (Kühlbrandt, 2014; wwPDB Consortium, 2024, 2025a).

The structure determination process of proteins involves several steps, from expression and purification, crystallisation or plunge-freezing, data collection, data processing (indexing, integration, scaling and merging in reciprocal space or image processing in real space), to model building and finally refinement. In this thesis, data originating from XRD, neutron and electron crystallography (ND and ED, respectively), as well as cryo-EM experiments have been utilised, either from published sources or kindly provided by collaborators.

As the aim of this thesis is to improve the accuracy and interpretation of the models explaining the data, the experimental setup will not be much further discussed. For the experimental aspects, the reader is referred to the literature, such as Bernhard Rupp's excellent book "Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology" (Rupp, 2009). In the case of cryo-EM, nobel laureate (and also considered to be the founder of single-particle cryo-EM) Joachim Frank's book "Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in

Their Native State" makes for a good read (Frank, 2005).

In the following chapter, I will provide a brief overview of the theoretical foundations upon which macromolecular structure determination rests.

## 2.1 Crystallography

### 2.1.1 Crystals

According to the International Union of Crystallography (IUCr), a solid is a crystal "if its atoms, ions and/or molecules form, on average, a long-range ordered arrangement" (real space) or "if it has essentially a sharp diffraction pattern" (reciprocal space), when shone upon with radiation with a wavelength on the order of Ångström (IUCr, 1992, 2021). The real-space definition implies that a crystal has some kind of periodic arrangement of a repeated unit, the unit cell. Mathematically, the unit cell can be described as the smallest possible translationally repeating unit in three dimensions, defined by its basis vectors $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$. Equivalently, the unit cell can also be defined by its lattice parameters: the lengths of the unit cell edges ($a_1$, $a_2$, $a_3$) and the angles between them. If the unit cell is repeated in all three dimensions, this forms the real space (or "direct") crystal lattice. Under the assumption that a crystal is perfect (i.e. the repetitions of the unit cell are identical), as well as infinitely periodic, this means that any local physical property $\rho(\mathbf{r})$, where $\mathbf{r} = x\,\mathbf{a}_1 + y\,\mathbf{a}_2 + z\,\mathbf{a}_3$ is a positional vector in fractional coordinates ($0 \leq x, y, z \leq 1$), of the crystal is invariant under any translation of the form $\mathbf{t} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3$, where $u_1$, $u_2$ and $u_3$ are integers. Put another way, $\rho(\mathbf{r})$ is a periodic function of $\mathbf{r}$, with periods $a_1$, $a_2$ and $a_3$ along the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, respectively:

$$\rho(\mathbf{r} + \mathbf{t}) = \rho(\mathbf{r}). \tag{2.1}$$

The reciprocal lattice can then be obtained by defining a new set of basis vectors, $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$, such that

$$\mathbf{b}_1 = \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3} \qquad \mathbf{b}_2 = \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3} \qquad \mathbf{b}_3 = \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}, \tag{2.2}$$

where $\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 = V$ is the volume of the unit cell. By definition, each vector in equation 2.2 is orthogonal to the corresponding real space lattice vectors, i.e. $\mathbf{b}_i \cdot \mathbf{a}_j = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta (($\delta_{ij} = 1$ if $i = j$, 0 if $i \neq j$)). A point in the reciprocal lattice is then given by the vector $\mathbf{h} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3$, where $h$, $k$ and $l$ are integers. The $(h, k, l)$ indices are usually referred to as "Miller indices".

Within the unit cell, additional symmetry elements can exist, such as mirror planes, rotational axes and inversion centers or any combination of them. In three dimensions, there are 230 possible unique space groups (combinations of symmetry elements). Due to the handedness of proteins (i.e. no mirror planes or inversion centers allowed), only 65 of these are permissible for describing their crystal structures. The smallest possible set of elements that under the symmetry operators from a specific space group fills the entire unit cell is called the asymmetric unit (Kittel, 2004; Rhodes, 2006; Rupp, 2009).

As it turns out, the periodicity of a crystal as given by equation 2.1 implies an ideal situation for Fourier analysis.

### 2.1.2 Diffraction

In 1912, Max von Laue and co-workers discovered that X-rays could be diffracted by crystals (for which he was later awarded the Nobel Prize in Physics in 1914) and published what would be known as the Laue equations, which give the necessary conditions for an incident wave to give rise to a diffraction pattern under elastic scattering (Laue, 1913, 1920).

In order to derive the Laue equations, we first introduce the concept of a wave vector, where the direction of the wave vector indicates the direction the wave is propagating, and the norm is taken to be $1/\lambda$, where $\lambda$ is the wavelength. A scattering event can then be described as the change in the wave vector between the incoming, $\mathbf{k}_i$, and outgoing, $\mathbf{k}_o$, wave vectors, i.e. $\Delta\mathbf{k} = \mathbf{s} = \mathbf{k}_o - \mathbf{k}_i$, where $\mathbf{s}$ is called the "scattering vector". For two volume elements of $\rho(\mathbf{r})$, the path difference for a wave emanating from a point $O$ (for simplicity, placed at the origin, see figure 2.1) and a point $P$ a distance $r = |\mathbf{r}|$ apart is then simply[1] the difference between the projection of $\mathbf{r}$ onto $\mathbf{k}_o$ and $\mathbf{k}_i$, respectively, and is given by

$$\Delta p = \lambda\mathbf{r}\cdot\mathbf{k}_o - \lambda\mathbf{r}\cdot\mathbf{k}_i = \lambda\mathbf{s}\cdot\mathbf{r}. \tag{2.3}$$

The phase difference in turn is obtained by multipling $\Delta p$ with $2\pi/\lambda$:

$$\Delta\varphi = 2\pi\mathbf{s}\cdot\mathbf{r}. \tag{2.4}$$

Owing to the superposition principle, maximal constructive interference between the waves is obtained when the phase difference is a multiple of $2\pi$, i.e. whenever $\mathbf{s}\cdot\mathbf{r} = n$, where $n$ is an integer. In particular, setting $\mathbf{r} = \mathbf{t}$ gives rise to the Laue equations:

$$\mathbf{s}\cdot\mathbf{a}_1 = n_1 \qquad \mathbf{s}\cdot\mathbf{a}_2 = n_2 \qquad \mathbf{s}\cdot\mathbf{a}_3 = n_3, \tag{2.5}$$

[1] Note that both the incoming and outgoing wave vectors are in units of $1/\lambda$ in the case of elastic scattering, i.e. $|\mathbf{k}_i| = |\mathbf{k}_o| = 1/\lambda$.

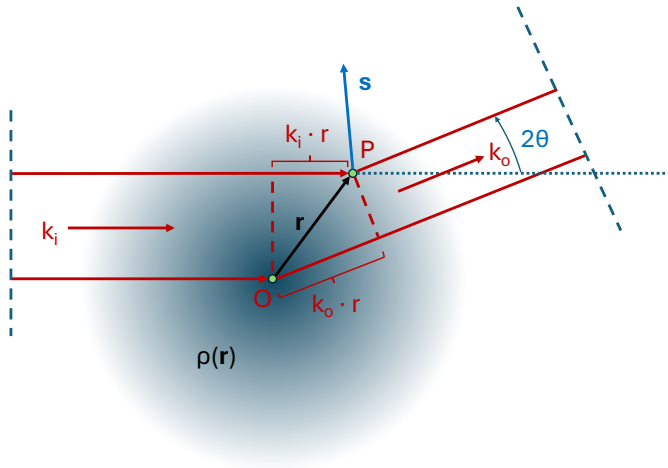



**Figure 2.1:** Schematic illustration of a scattering event for two volume elements of $\rho(\mathbf{r})$, separated by a distance $r = |\mathbf{r}|$, for an incoming wave vector $\mathbf{k}_i$ and an outgoing wave vector $\mathbf{k}_o$. Figure adapted from (Rupp, 2009).

where $n_i$ are integers. The Laue equations thus describe the conditions under which (maximal) constructive interference occurs for diffraction from a crystal lattice. In order to see that these conditions are satisfied for points in the reciprocal lattice, the translational invariance from equation 2.1 can be expanded in a Fourier series:

$$\rho(\mathbf{r}) = \sum_{\mathbf{h}} \rho_{\mathbf{h}} e^{i2\pi\mathbf{h}\cdot\mathbf{r}} = \sum_{\mathbf{h}} \rho_{\mathbf{h}} e^{i2\pi\mathbf{h}\cdot(\mathbf{r}+\mathbf{t})} = \rho(\mathbf{r}+\mathbf{t}) \tag{2.6}$$

which holds true if and only if $e^{i2\pi\mathbf{h}\cdot\mathbf{t}} = 1$, or, equivalently, that $\mathbf{h}\cdot\mathbf{t} \in \mathbb{Z}$. Thus, the Laue equations are satisfied for points in the reciprocal lattice whenever $\mathbf{s} = \mathbf{h}$. This also shows that the reciprocal lattice is the Fourier transform of the real-space lattice.

For a scatterer $j$ positioned at $\mathbf{r}_j = x_j\mathbf{a}_1 + y_j\mathbf{a}_2 + z_j\mathbf{a}_3$, the Laue equations also implies that $\mathbf{s}\cdot\mathbf{r}_j = hx_j + ky_j + lz_j = \mathbf{h}\cdot\mathbf{x}_j$.

From the Laue equations, it is also possible to derive the well-known Bragg's law[2], which gives the condition for constructive interference from two lattice planes with Miller indices $(h, k, l)$, separated by a distance, $d_{hkl}$, as

$$n\lambda = 2d_{hkl}\sin\theta, \tag{2.7}$$

where $\theta$ is the angle between the incident wave and the lattice planes and $n$ is an integer. Interestingly, Bragg's law implies a maximum for the separation of two lattice planes

[2]This derivation is left as an exercise for the reader. Hint: the distance between lattice planes is given by $d_{hkl} = \frac{n}{|\mathbf{h}|}$, $n \in \mathbb{N}^+$.

(i.e. the maximum resolution, where "resolution" is defined as the ability to distinguish between two closely spaced objects) that can be resolved at a wavelength $\lambda$ in the form of $d_{hkl}^{min} = \lambda/2$, for a scattering angle $\theta = \pi/2$. In practise, the maximum resolution is given by $d_{hkl}^{min} = \lambda/(2\sin\theta_{max})$, where $\theta_{max}$ is the maximum scattering angle that can be measured in the experiment.

Equation 2.7 can also be argued to hold true from simpler geometrical arguments, as was done by father and son Bragg (Bragg and Bragg, 1913).

### 2.1.3 Atomic scattering factors

So far, the scattering event has been considered for two distinct volume elements of $\rho(\mathbf{r})$. The total scattered wave from a single scatterer $j$, assumed centered at $\mathbf{r} = 0$, is obtained by integrating the density of this scatterer, $\rho_j(\mathbf{r})$, with a relative phase $\varphi_j = 2\pi\mathbf{s}\cdot\mathbf{r}$ for each volume element, over the volume of the scatterer:

$$f_j = \int \rho_j(\mathbf{r})\, e^{i2\pi\mathbf{s}\cdot\mathbf{r}}\, \mathrm{d}\mathbf{r}, \tag{2.8}$$

where $f_j$ is called the "atomic scattering factor" or "atomic form factor".

The atomic form factor is a measure of the scattering strength of scatterer $j$. In general, at least when it comes to proteins, it often is assumed that scatterers of the same constitution, i.e. for example the same element or isotope, have the same atomic form factor, and that the atomic form factor is independent of the chemical environment of the scatterer. Furthermore, the atomic form factor is often assumed to be spherically symmetric. Together, this is known as the independent atom model (IAM) (Compton, 1915).

In the case of XRD, the scattering event is due to photons interacting with the electron density surrounding the nucleus. Under the above assumptions for the IAM, this density does not change between scatterers of the same type and can thus be precalculated and tabulated. This has been done in the form of the Cromer–Mann coefficients, where nine-parameter Gaussian approximations have been fitted to spherically averaged Hartree–Fock wave functions for different elements and ions thereof (Cromer and Mann, 1968; Brown et al., 2006; Grabowsky et al., 2020). While this allows for efficient lookup tables in crystallographic software, for (very) high-resolution structures, the shortcomings of the IAM do become apparent. It can also be noted from equation 2.8 that in the case of XRD $\lim_{|\mathbf{s}|\to 0} f_j^X = Z_j$, where $Z_j$ is the number of electrons of scatterer $j$. That is, in the forward scattering direction, the scattering strength is proportional to the number of electrons for that scatterer. There will also be a strong dependence on the scattering angle (or equivalently, $|\mathbf{s}|$). On the other hand, there is only a small difference in the scattering strength

for elements next to each other in the periodic table (e.g. carbon, nitrogen and oxygen will all scatter similarly), which can make it difficult to distinguish between them. Likewise, hydrogen, which only has one electron, has a very weak scattering strength with regards to XRD. Hydrogen atoms are thus often not visible in XRD structures, unless the data is of very high resolution (better than $\approx 1.2$ Å).

For ND, the scattering event is due to neutrons interacting with the nucleus of the atom. Owing to the fact that the nuclei of atoms are very small (on the order of $\approx 10^{-5}$ Å) compared to the wavelengths of the neutrons used in diffraction experiments (on the order of $\approx 1$ Å), the scattering event is in practice independent of the scattering angle and the nuclei act as point scatterers. In turn, $\rho_j(\mathbf{r})$ in equation 2.8 can be approximated as a Dirac delta function multiplied by a constant, i.e. $\rho_j(\mathbf{r}) \approx \bar{b}_j \delta(\mathbf{r})$, resulting in

$$f_j^N \approx \int \bar{b}_j \delta(\mathbf{r}) e^{i2\pi \mathbf{s}\cdot\mathbf{r}} \, \mathrm{d}\mathbf{r} = \bar{b}_j, \tag{2.9}$$

where $\bar{b}_j$ is the "neutron scattering length" of scatterer $j$. Neutron scattering lengths cannot be be calculated from nuclear theory, but must be determined experimentally (Varley, 1992). Furthermore, the scattering length depends on the specific isotope and can even be negative. Interestingly, hydrogen has a negative scattering length ($\bar{b}_H = -3.74$ fm), while deuterium has a positive one ($\bar{b}_D = 6.67$ fm), which can be exploited in ND experiments through deuteration of the sample, in order to make the positions of hydrogen atoms more visible (Varley, 1992; Shu et al., 2000; Blakeley et al., 2008; Schröder and Meilleur, 2021).

With ED, the scattering event is due to electrons interacting with the electrostatic potential distribution of the atom, which has contributions from the positively charged nucleus and the surrounding negatively charged electron density. The electron form factor can be calculated from the corresponding XRD atomic form factor, $f_j^X$, through the Mott–Bethe formula:

$$f_j^{ED} = \frac{m_0 e^2}{8\pi^3 \hbar^2 \varepsilon_0} \left( \frac{Z_j - f_j^X}{|\mathbf{s}|^2} \right), \tag{2.10}$$

where $m_0$ is the electron rest mass, $e$ is the elementary charge, $\hbar$ is the reduced Planck constant, $\varepsilon_0$ is the permittivity of free space and $|\mathbf{s}| = 2\sin\theta/\lambda$ is the magnitude of the scattering vector[3], with $\theta$ being half the scattering angle and $\lambda$ the electron wavelength (Bethe, 1930; Mott, 1930). Notably, the electron scattering factor has a much stronger dependence on the charge of the scatterer than the corresponding X-ray scattering factor, which implies that it, at least in theory, should be possible to probe the charge distribution of a system with ED (Peng, 1998; Lobato and Van Dyck, 2014; Saha et al., 2022; Pacoste,

[3]If the angle between the incoming, $\mathbf{k}_i$, and outgoing, $\mathbf{k}_o$, wave vectors is $2\theta$, with $|\mathbf{k}_i| = |\mathbf{k}_o| = 1/\lambda$, then $|\mathbf{s}|^2 = \mathbf{s}\cdot\mathbf{s} = (\mathbf{k}_o - \mathbf{k}_i)\cdot(\mathbf{k}_o - \mathbf{k}_i) = |\mathbf{k}_o|^2 - 2|\mathbf{k}_o||\mathbf{k}_i|\cos(2\theta) + |\mathbf{k}_i|^2 = 2(1 - 2\cos(2\theta))/\lambda^2 = 4\sin^2\theta/\lambda^2 \implies |\mathbf{s}| = 2\sin\theta/\lambda$.

2025).

### 2.1.4 Structure factors

The total scattering function for a unit cell for a specific reflection $(h, k, l)$, $F_{hkl}$, is the sum of the contributions from all scatterers in the unit cell to that reflection, each with a relative phase $\varphi_j = 2\pi\mathbf{s} \cdot \mathbf{r}_j$. For a unit cell with $N$ scatterers, each with their own atomic form factor $f_j$ and position $\mathbf{r}_j$, this function is given by

$$F_{hkl} = \sum_{j=1}^{N} f_j \, e^{i2\pi\mathbf{s}\cdot\mathbf{r}_j} = \sum_{j=1}^{N} f_j \, e^{i2\pi\mathbf{h}\cdot\mathbf{x}_j} \tag{2.11}$$

and is called the "structure factor". With a crystal consisting of $M$ unit cells, the total scattering function is then simply $F = MF_{hkl}$ due to the periodicity of the crystal, which also reveals that the crystal acts as an amplifier. As equation 2.11 shows, the structure factor can be calculated for a given model of scatterers in the unit cell, which is why the notation $F_{hkl}^{calc}$ is often used for calculated structure factors. Conversely, if the structure factors are known from experiment, these are often denoted $F_{hkl}^{obs}$.

The total density, $\rho(\mathbf{r})$, in the unit cell that is causing the scattering can be interpreted as a sum of the respective densities of all scatterers in the unit cell, i.e.

$$\rho(\mathbf{r}) = \sum_{j=1}^{N} \rho_j(\mathbf{r} - \mathbf{r}_j), \tag{2.12}$$

where $\rho_j(\mathbf{r} - \mathbf{r}_j)$ is the contribution to the total density of scatterer $j$ at $\mathbf{r}$. Combining[4] equations 2.8, 2.11 and 2.12 gives that

$$F_{hkl} = \int \rho(\mathbf{r}) e^{i2\pi\mathbf{s}\cdot\mathbf{r}} \, \mathrm{d}\mathbf{r} = \int \rho(\mathbf{r}) e^{i2\pi\mathbf{h}\cdot\mathbf{x}} \, \mathrm{d}\mathbf{r} = \mathcal{F}\{\rho(\mathbf{r})\}, \tag{2.13}$$

which shows that the structure factors are Fourier transforms of the density $\rho(\mathbf{r})$. The density can in turn be obtained as the inverse Fourier transform of the structure factors:

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_{hkl} F_{hkl} \, e^{-i2\pi\mathbf{h}\cdot\mathbf{x}} = \frac{1}{V}\sum_{hkl} |F_{hkl}| e^{i\varphi_{hkl}} e^{-i2\pi\mathbf{h}\cdot\mathbf{x}} = \frac{1}{V}\sum_{hkl} |F_{hkl}| e^{-i(2\pi\mathbf{h}\cdot\mathbf{x} - \varphi_{hkl})}, \tag{2.14}$$

[4]Note that $\mathbf{r}$ in equation 2.8 now is assumed to be $\mathbf{r} - \mathbf{r}_j$ and that $\mathbf{r}_j$ is somewhere in the unit cell, not necessarily at the origin.

where $V$ is the volume of the unit cell and $\varphi_{hkl}$ is the phase of the structure factor. Equation 2.14 is of fundamental importance in crystallography, as it allows for the calculation of the density, $\rho(\mathbf{r})$, from the structure factors, $F_{hkl}$ (Rhodes, 2006; Rupp, 2009).

The structure factor is typically a complex number, but the scattered intensity, $I_{hkl}$, which is what is measured by the detector in a diffraction experiment, is proportional to the square of the norm of the structure factor, i.e.

$$I_{hkl} \propto |F_{hkl}^{obs}|^2 = \overline{F_{hkl}^{obs}} F_{hkl}^{obs}, \tag{2.15}$$

which is real-valued. Unfortunately, this also means that the phase information is lost during the measurement. This is known as the "phase problem" in crystallography.

### 2.1.5 Atomic displacement and occupancy

In a real crystal, the atoms are not static but vibrate around their equilibrium positions, as well as that the unit cells may not be identical to each other, due to defects and disorder. A more correct description of the density causing the scattering events is thus a time- and space-averaged density, i.e.

$$\langle \rho(\mathbf{r}) \rangle \approx \sum_{j=1}^{N} n_j \int \rho_j(\mathbf{r} - \mathbf{r}_j) P_j(\mathbf{u}_j) \, \mathrm{d}\mathbf{r}, \tag{2.16}$$

where $n_j$ is the occupancy of scatterer $j$ and $P_j(\mathbf{u}_j)$ is the probability distribution of the displacement vector $\mathbf{u}_j$ of scatterer $j$ from its over the unit cells averaged equilibrium position, $\mathbf{r}_j$ (Trueblood et al., 1996).

The displacement probabilities can be accounted for by introducing atomic displacement parameters (ADPs) through the Debye–Waller factor, which under the assumption of harmonic oscillations around the scatterers equilibrium positions (which implies a Gaussian distribution of the atomic displacements) is given by

$$T_j(\mathbf{h}) = e^{-2\pi^2 \langle (\mathbf{s} \cdot \mathbf{u}_j)^2 \rangle}, \tag{2.17}$$

where $\langle (\mathbf{s} \cdot \mathbf{u}_j)^2 \rangle$ is the mean square displacement of scatterer $j$ in the direction of the scattering vector $\mathbf{s}$ (Debye, 1913; Waller, 1923; Trueblood et al., 1996).

Assuming isotropic movement, equation 2.17 simplifies to

$$T_j^{iso}(\mathbf{h}) = e^{-8\pi^2 \langle u_j^2 \rangle (\sin\theta/\lambda)^2} = e^{-B_j (\sin\theta/\lambda)^2}, \tag{2.18}$$

where $B_j = 8\pi^2 \langle u_j^2 \rangle$.

In the case of anisotropic movement, six new parameters per scatterer are instead introduced and equation 2.17 becomes

$$T_j^{ani}(\mathbf{h}) = e^{-2\pi^2 \mathbf{h}^T \mathbf{U}_j \mathbf{h}}, \tag{2.19}$$

where $\mathbf{U}_j$ is a $3 \times 3$ symmetric tensor (Trueblood et al., 1996; Rupp, 2009; Hoser and Madsen, 2025).

The ADPs and occupancies are readily included in the structure factor by augmenting the atomic form factor with the Debye–Waller factor and the corresponding occupancy, i.e.

$$f_j \to n_j f_j T_j(\mathbf{h}). \tag{2.20}$$

Typically, isotropic ADPs are used for lower-resolution structures, whereas anisotropic ADPs are used for higher-resolution structures (better than $\approx 1.5$ Å), where the risk of overfitting is lower.

### 2.1.6 Bulk solvent

In macromolecular crystallography, the crystal is not only made up of the macromolecule, but it also contains a significant amount of solvent, typically water. This bulk solvent is present in the crystal lattice voids and constitutes on average about 50% of the unit cell volume and is generally disordered (Matthews, 1968; Chruszcz et al., 2008; Weichenberger et al., 2015). While disordered, the bulk solvent still contributes to the measured intensities and must thus be accounted for in the structure factor calculations, especially at low resolutions.

The most common approach to model the bulk solvent contribution is to use a binary solvent mask, which is zero inside the protein and for well-resolved solvent molecules, whereas it is one in the disordered solvent region. The mask can be calculated through use of Babinet's principle or exact asymmetric units (Moews and Kretsinger, 1975; Jiang and Brünger, 1994; Fenn et al., 2010; Grosse-Kunstleve et al., 2011; Afonine et al., 2013). An additional structure factor is then calculated as the Fourier transform of the solvent mask, $F_{hkl}^{mask}$, which is added to $F_{hkl}^{calc}$ to obtain the total structure factor. Additionally, scaling factors are often included to account for the overall scaling of the data and the solvent contribution, i.e.

$$F_{hkl}^{model} = k_{total} \left( F_{hkl}^{calc} + k_{mask} F_{hkl}^{mask} \right), \tag{2.21}$$

where $k_{total}$ and $k_{mask}$ are scaling factors for the total structure factor and the solvent mask contribution, respectively, with $k_{mask}$ given by

$$k_{mask} = k_{sol} e^{(-B_{sol}|\mathbf{s}|^2/4)}, \tag{2.22}$$

where $k_{sol}$ and $B_{sol}$ are the flat bulk-solvent model parameters (Afonine et al., 2005, 2013, 2023).

Other methods include for example modelling the bulk solvent using an exponential decay function (Moews and Kretsinger, 1975; Tronrud, 1997). An interesting recent development is modelling the bulk solvent molecules explicitly, through use of molecular dynamics (MD) simulations (Mikhailovskii et al., 2024).

### 2.1.7 The phase problem

Ultimately, the goal of a crystallographic experiment is to obtain the density $\rho(\mathbf{r})$, into which a model can be built which represents the structure in the crystal. Equation 2.14 provides a way to calculate the density from the structure factors, but requires knowledge of both the amplitudes, $|F_{hkl}|$, and the phases, $\varphi_{hkl}$, where the experiment provides only the amplitudes through the measured intensities. From the right-hand side of equation 2.11, it is also clear that a model of what is causing the diffraction pattern carries information about the phases, as the positions, $\mathbf{x}_j$, of the scatterers are included in the exponential term.

In order to get started with the structure determination process, initial phases must thus be obtained through other means. Options include, but are not limited to (Taylor, 2010):

- **Multiple isomorphous replacement (MIR)**, where heavy atoms are introduced into the crystal, which cause changes in the intensities of the diffracted beams. By comparing the intensities of the native crystal with those of the derivatised crystal, initial positions of the heavy atoms can be determined and in turn initial phases for all reflections estimated, using Patterson methods (Patterson, 1934; Green et al., 1954; Rupp, 2009). Ideally, the heavy atoms should not alter the overall structure of the macromolecule or the unit cell, hence the term "isomorphous". In order to resolve phase ambiguities, data from multiple different heavy-atom derivatives are typically required (Taylor, 2003; Cowtan, 2003).
- **Multi- or single-wavelength anomalous diffraction (MAD or SAD)**, where the scattering strength of certain atoms (e.g. selenium, iodine or metals) varies as a function of the wavelength of the incoming X-rays, when the wavelength is tuned to near an absorption edge of that element. By collecting data at several wavelengths (MAD), or just one wavelength (SAD), initial positions of these anomalous scatterers can be

determined and thus initial phases for all reflections estimated, in a manner similar to MIR (Hendrickson et al., 1990; Hendrickson, 1991; Cowtan, 2003; Rhodes, 2006).

- **Molecular replacement** (**MR**), where a previously determined structure of a similar macromolecule is used as a starting point to estimate the phases, typically either through the wealth of previously solved structures available in the PDB and EMDB or through machine learning methods (Rossmann and Blow, 1962; Rossmann, 1990; Jumper et al., 2021; McCoy et al., 2022; Barbarin-Bocahu and Graille, 2022; Yang et al., 2023; Terwilliger et al., 2024).

Once an initial set of positions for some atoms and thereby initial estimates of the structure factor phases has been obtained, an initial density can be calculated through equation 2.14, into which more of the model can be built. This improved model is then used to calculate new phases, which in turn can be used to calculate a new density, into which more of the model can be built, and so on, and the structure is then improved in an iterative manner.

## 2.2 Cryogenic electron microscopy

Contrary to X-rays and neutrons, electrons are charged particles and are thus subject to the Lorentz force (i.e. $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, where $q$ is the charge of the particle, $\mathbf{E}$ is the electric field, $\mathbf{v}$ is the velocity of the particle, $\mathbf{B}$ is the magnetic field and $\mathbf{F}$ is the resulting force acting on the particle) when moving in a magnetic field. In turn, electrons can hence be focused using electromagnetic lenses, which is the basis of the electron microscope. In electron microscopy (EM), a beam of electrons is accelerated to high energies, typically 100–300 keV. The electrons are then scattered by the sample and refocused through electromagnetic lenses to form magnified images on a detector. From the de Broglie relation ($\lambda = h/p$, where $h$ is the Planck constant and $p$ is the momentum of the electron), the corresponding wavelengths of the electrons at these energies are on the order of 0.04–0.02 Å (de Broglie, 1924).

In theory, this means that EM with ease can provide atomic resolution real space representations of the sample. However, in practise, the resolution is limited by a number of factors, such as lens aberrations, sample quality and, in particular for biological samples, radiation damage (Baker and Henderson, 2012; Milne et al., 2013; Thompson et al., 2016).

Two main types of electron microscopy exist: transmission electron microscopy (TEM) and scanning electron microscopy (SEM). In TEM, a beam of electrons is transmitted through a thin sample, while in SEM the surface of a sample is scanned with a focused beam of electrons. For biological samples, TEM is typically used, while SEM is more commonly used for materials science applications. In order to obtain high-resolution images

of macromolecules, TEM is typically used in cryogenic mode in order to prevent radiation damage, where the sample is rapidly frozen in vitreous ice to preserve its native structure and to obtain images from as many different orientations as possible (Milne et al., 2013). This particular technique is often referred to as "cryo-EM". The techniques used to rebuild the three-dimensional structure are often grouped together under the umbrella "single-particle analysis" (SPA), as the images are of individual particles. A more apt abbreviation for the experimental setting used for macromolecular structure determination in cryo-EM is perhaps then "cryo-TEM-SPA", but the term "cryo-EM" has become the most commonly used one in the literature.

Compared to reciprocal-space methods, cryo-EM has the advantage of not requiring crystallisation of the sample, which can be a major bottleneck in structure determination. Especially membrane proteins are notoriously difficult to crystallise (Carpenter et al., 2008; Loll, 2014). On the other hand, cryo-EM typically requires large amounts of data and extensive computational resources for image processing and reconstruction. However, recent advances in detector technology, image processing algorithms and machine learning have significantly improved the capabilities of cryo-EM, making it a powerful tool for structural biology (Saibil, 2022; Chua et al., 2022; Vilas et al., 2022).

### 2.2.1 Image formation

The primary requirement for observing structure in images is the presence of contrast. Typically, two cases of contrast in TEM are considered: amplitude and phase contrast. Amplitude contrast arises when the amplitude of the incoming wave is changed due to absorption or loss of electrons through wide scattering by the sample, while phase contrast arises when the phase of the incoming wave is changed due to the electrostatic potential of the sample (i.e. elastic scattering). In TEM, phase contrast is the dominant form of contrast, as biological samples are weak phase objects, i.e. they are thin and do not absorb or scatter a significant amount of the incoming electrons, but rather change the phase of the incoming electron wave. This is commonly known as the "weak phase object approximation" (WPOA) (Reimer and Kohl, 2008; Milne et al., 2013; CryoSPARC, 2025).

In order to understand how images are formed in TEM, we can start by considering the interaction between the incoming electron wave and the sample. The incoming electron wave can be described as a plane wave,

$$\psi_i(\mathbf{r}) = Ae^{i\mathbf{k}\cdot\mathbf{r}}, \tag{2.23}$$

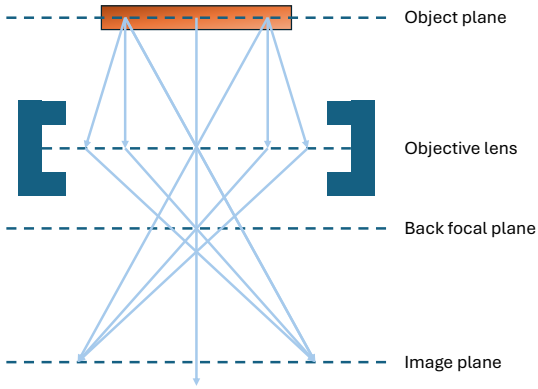



Figure 2.2: Schematic illustration of image formation under phase contrast in cryo-EM. An incoming electron wave interacts with the sample (placed in the object plane), resulting in an exit wave that has experienced a phase shift due to the electrostatic potential of the sample. The exit wave is then focused by the objective lens, which introduces another phase shift due to lens aberrations. The wave in the back focal plane of the objective lens is given by the Fourier transform of the exit wave multiplied by the phase shift introduced by the lens. Finally, the image formed on the detector, positioned at the image plane, is given by the inverse Fourier transform of the wave in the back focal plane, and what is observed on the detector is the intensity of the wave.

where $A$ is the amplitude of the wave, $\mathbf{k}$ is the wave vector and $\mathbf{r}$ is the position vector[5]. When the electron wave interacts with the sample, it experiences a phase shift due to the electrostatic potential, $\phi(\mathbf{r})$, of the sample. Under the WPOA, the wave after passing through the sample can then be approximated as

$$\psi_e(\mathbf{r}) = \psi_i(\mathbf{r}) e^{i\phi(\mathbf{r})} \approx A e^{i\mathbf{k}\cdot\mathbf{r}}(1 + i\Phi(x, y)), \tag{2.24}$$

where $\Phi(x, y) = \int \phi(\mathbf{r})\, \mathrm{d}z$ (Frank, 2005). The exit wave is then focused by the objective lens of the microscope, which introduces another phase shift, $\chi(\mathbf{q})$, due to lens aberrations and defocus. The wave in the back focal plane of the objective lens is then given by the Fourier transform of the exit wave multiplied by this phase shift from the lens:

$$\psi_{bf}(\mathbf{q}) = \mathcal{F}\{\psi_e(\mathbf{r})\} e^{i\chi(\mathbf{q})}. \tag{2.25}$$

Notably, if the sample is a crystal and the WPOA holds, a diffraction pattern of the crystal can thus be observed in the back focal plane[6], which is exactly the idea behind electron crystallography (Frank, 2005; Reimer and Kohl, 2008).

[5]Which in the case of real space is not limited to the unit cell as in crystallography, but can take on any value in $\mathbb{R}^3$.

[6]As intensities are what can be observed, the phase shift introduced by the lens does not affect the diffraction pattern, as $|e^{i\chi(\mathbf{q})}|^2 = 1$.

Finally, the image formed on the detector in the image plane of the objective lens is given by the inverse Fourier transform of the wave in the back focal plane, i.e.

$$\psi_{img}(\mathbf{r}) = \mathcal{F}^{-1}\{\psi_{bf}(\mathbf{q})\}. \tag{2.26}$$

Like in the case of crystallography, what is observed on the detector is the intensity of the wave, which is given by

$$I(x, y) \propto |\psi_{img}(\mathbf{r})|^2 = \overline{\psi_{img}(\mathbf{r})}\psi_{img}(\mathbf{r}). \tag{2.27}$$

### 2.2.2 The contrast transfer function

The sine of the phase shift introduced by the lens is known as the "contrast transfer function" (CTF), which describes how different spatial frequencies are transferred to the image. Several models for the CTF exist, but most follow the same basic form:

$$\mathrm{CTF}(\mathbf{q}) = \sin\left(\frac{\pi}{2}C_s\lambda^3|\mathbf{q}|^4 - \pi\Delta z\lambda|\mathbf{q}|^2\right), \tag{2.28}$$

where $C_s$ is the spherical aberration coefficient, $\Delta z$ is the defocus, $\lambda$ is the wavelength of the electrons and $\mathbf{q}$ is the spatial frequency vector (Frank, 2005). From equation 2.28, it is clear that the CTF depends on the spherical aberration coefficient, which is a property of the microscope, as well as the defocus, which can be controlled during the experiment. Due to the sine function, the CTF oscillates between positive and negative values and will for some spatial frequencies be zero, meaning that those spatial frequencies are not transferred to the image. Shifting the defocus will shift the CTF and thus make it possible to recover the lost information by taking images at different defocus values. Additionally, if a sample is perfectly in focus (i.e. $\Delta z = 0$), for low spatial frequencies the CTF will be close to zero, meaning that low-resolution information is not transferred to the image. Because of these reasons, cryo-EM images are typically taken with some defocus, at different values, in order to improve the contrast (Frank, 2005; CryoSPARC, 2025).

In addition to the CTF, the spatial and temporal coherence of the electron beam also affects the transfer of spatial frequencies to the image. These effects can be included in the CTF as multiplicative factors, resulting in an effective CTF given by

$$\mathrm{CTF}_{\mathit{eff}}(\mathbf{q}) = \mathrm{CTF}(\mathbf{q})E_t(\mathbf{q})E_s(\mathbf{q}), \tag{2.29}$$

where $E_t(\mathbf{q})$ and $E_s(\mathbf{q})$ are temporal and spatial envelope functions, respectively (Frank, 2005; Reimer and Kohl, 2008). Obtaining a good estimate of the CTF is crucial for high-resolution structure determination in cryo-EM. While the spherical aberration coefficient, $C_s$, ideally is constant for a given microscope, the micrographed particles are frozen at various depths in the vitreous ice. In turn, this means that the defocus will vary slightly for each particle and as a consequence, the CTF should be estimated individually for each particle (Vilas et al., 2022).

### 2.2.3 Image processing and 3D reconstruction

A large number of images (often referred to as "micrographs") are then collected, each containing many individual particles, in as many different orientations as possible. Modern detectors record image stacks ("movies"), instead of only single images (Cheng et al., 2015). The goal of the image processing step is to extract the individual particles from the micrographs, align them to a common reference frame and then reconstruct a three-dimensional structure from the aligned particles. This process involves several steps, including (Vilas et al., 2022):

- **Motion correction**, where the individual frames of a movie are aligned to correct for beam-induced motion of the sample during the exposure.
- **CTF estimation**, where the CTF parameters are estimated from the micrographs.
- **Particle picking**, where individual particles are identified and extracted from the micrographs.
- **2D classification**, where the extracted particles are grouped into classes based on their similarity in 2D projections.

The signal-to-noise ratio (SNR) of the individual particles is typically very low, due to the low electron dose used in order to minimise radiation damage to the sample. However, by averaging together many particles, the SNR can be improved (Milne et al., 2013; Cheng et al., 2015).

With class averages from the 2D classification step available, the problem is then to reconstruct the three-dimensional electrostatic potential of the macromolecule from its two-dimensional projections, i.e. for a dataset containing $N$ class averages, $\{I_1, I_2, \ldots, I_N\}$, on a Cartesian grid $(x, y)$, the goal is to find $\phi(\mathbf{r})$ such that

$$I_i(x, y) = H_i * \int_{-\infty}^{\infty} \phi(R_i^T \mathbf{r})\, \mathrm{d}z + \text{“noise”}, \quad i = 1, 2, \ldots, N, \tag{2.30}$$

where $H_i$ is the point spread function (PSF) of the class, which is the Fourier transform of the CTF, and $R_i \in SO(3)$ is a rotation matrix which is unknown *a priori*. Estimation of $\phi(\mathbf{r})$ is called the "cryo-EM reconstruction problem" (Singer, 2018).

Several methods exist for solving the cryo-EM reconstruction problem. One is the common lines method, which relies on the fact that the Fourier transforms of any two two-dimensional projections of a three-dimensional object intersect along a common line (which is also known as the "Fourier slice theorem") (Van Heel, 1987; Cheng et al., 2015; Vilas et al.,

2022). By identifying these common lines between pairs of projections, the relative orientations of the projections can be determined and, additionally, a three-dimensional Fourier space can be assembled. The inverse Fourier transform of this three-dimensional Fourier space then gives the three-dimensional electrostatic potential (Frank, 2005; Sigworth, 2016).

More modern approaches rely on maximum likelihood estimation, which are implemented in software packages such as `RELION` and `CryoSPARC` (Scheres, 2012; Punjani et al., 2017). These methods attempt to solve the cryo-EM reconstruction problem probabilistically, where the goal is to find the electrostatic potential that maximises the likelihood of observing the experimental images. Additionally, machine learning approaches are fast gaining traction in cryo-EM reconstruction (Vilas et al., 2022).

Like in the 2D classification step, 3D classification can also be performed, where the particles are grouped into different classes based on their similarity in terms of their three-dimensional structure, allowing for the identification of different conformational states of the macromolecule (Scheres, 2012; Punjani et al., 2017; Vilas et al., 2022).

## 2.3 Refinement

Once a density or electrostatic potential map is available, a model of the macromolecule can be built into this density or map. The model is then subsequently refined in order to improve its fit, where according to the IUCr, refinement is defined as "the process of adjusting the parameters of a model to find values *most nearly* compatible with the observations" (IUCr, 1996). Ideally, the data alone should be sufficient to determine the model. In practise, this is often not the case, especially not for lower-resolution data, which is typically obtained for macromolecular structures. In order to improve the data-to-parameter ratio, prior knowledge about the system is therefore often included in the refinement process through restraints. The refinement target function, $T$, is then given by

$$T = w_{data} T_{data} + T_{restraints}, \tag{2.31}$$

where $T_{data}$ is a term describing the fit of the model to the experimental data, $T_{restraints}$ is a term describing the fit of the model to prior knowledge about the system and $w_{data}$ is a weight factor balancing the two terms. The goal of refinement is then to find the model parameters (i.e. coordinates, ADPs and occupancies) that minimises the target function $T$ (Tronrud, 2004; Urzhumtsev and Lunin, 2019).

The data term, $T_{data}$, depends on the type of experimental data used. In crystallography,

an older common choice for $T_{data}$ is a least-squares target function,

$$T_{data} = \sum_{hkl} w_{hkl} \left( |F_{hkl}^{obs}| - |F_{hkl}^{model}| \right)^2, \tag{2.32}$$

where $|F_{hkl}^{obs}|$ are the observed structure factor amplitudes, $|F_{hkl}^{model}|$ are the structure factor amplitudes calculated from the model through equation 2.21 and $w_{hkl}$ are weights, often chosen as $1/\sigma_{hkl}^2$, where $\sigma_{hkl}^2$ are the variances of the observed structure factor amplitudes (Arnold and Rossmann, 1988).

Equation 2.31 can also be interpreted in a Bayesian framework, where the goal is to find the model parameters that maximises the posterior probability, $P(\text{model}|\text{data})$, i.e.

$$P(\text{model}|\text{data}) \propto P(\text{data}|\text{model})P(\text{model}), \tag{2.33}$$

where $P(\text{data}|\text{model})$ is the likelihood of the data given the model and $P(\text{model})$ is the prior probability of the model (Sivia, 2006). Taking the logarithm of equation 2.33 and multiplying by $-1$ gives

$$-\log(P(\text{model}|\text{data})) \propto -\log(P(\text{data}|\text{model})) - \log(P(\text{model})), \tag{2.34}$$

which has the same functional form as equation 2.31, where $T_{data}$ corresponds to the negative log-likelihood and $T_{restraints}$ corresponds to the negative log-prior, i.e.

$$T_{data} = -\log(P(\text{data}|\text{model})), \quad T_{restraints} = -\log(P(\text{model})). \tag{2.35}$$

This means that minimising the total target $T$ in equation 2.31 is equivalent to maximising the posterior probability in equation 2.33, i.e. finding the most probable model given the data and prior knowledge (Murshudov et al., 1997; Pannu et al., 1998; Tronrud, 2004; McCoy, 2004).

Under an assumption[7] that the individual experimental observations, $d_i$, are independent conditional on the model parameters, as well as that the individual parameters of the model, $r_j$, are mutually independent *a priori*, the total target $T$, for $N$ observations and $M$ interactions, becomes

$$T = \sum_{i=1}^{N} -\log(P(d_i|\text{model})) + \sum_{j=1}^{M} -\log(P(r_j)). \tag{2.36}$$

[7]This is a common assumption in Bayesian statistics, particularly in the context of parameter estimation. Unfortunately, this is typically not a very good approximation for the restraints term in the case of the usual components used, as for example bond lengths and angles are often correlated. Nor is it a good approximation for the experimental observations of the structure factors, as they are often not independent, although this correlation tends to be small. For these reasons, a weight factor, $w_{data}$, is introduced in the refinement process.

If the prior probability of the model is assumed to be a multivariate Gaussian distribution (which implies a multivariate harmonic potential describing the model), the restraints term is then given by

$$T_{restraints} \propto \sum_i \frac{(x_i - \bar{x}_i)^2}{\sigma_{x_i}^2}, \tag{2.37}$$

where $x_i$ are the model parameters, $\bar{x}_i$ are the target values for the parameters and $\sigma_{x_i}^2$ are the variances of the parameters (McCoy, 2004).

Furthermore, the likelihood in a crystallographic setting, under a Gaussian error model, can be written as

$$P\left(F_{hkl}^{obs}|F_{hkl}^{model}\right) = \begin{cases} \frac{1}{\pi\sigma_\Delta^2} \exp\left(-\frac{|F_{hkl}^{obs} - DF_{hkl}^{model}|^2}{\sigma_\Delta^2}\right), & \text{acentric reflections,} \\ \frac{1}{(2\pi\sigma_\Delta^2)^{1/2}} \exp\left(-\frac{|F_{hkl}^{obs} - DF_{hkl}^{model}|^2}{2\sigma_\Delta^2}\right), & \text{centric reflections,} \end{cases} \tag{2.38}$$

where $D$ is the Luzatti factor ($D = \langle \cos(2\pi\Delta\mathbf{r}_j \cdot \mathbf{s}) \rangle$) and $\sigma_\Delta^2$ is the variance of the observed structure factors (Luzzati, 1952; Read, 1990; Bricogne, 1997; Pannu et al., 1998; Read, 2001; McCoy, 2004; Rupp, 2009). Acentric reflections are reflections where no symmetry operation in the space group exists that sends the reflection $(h, k, l)$ to $(-h, -k, -l)$ and centric reflections are reflections where such a symmetry operation exists.

The likelihood in equation 2.38 can in turn be used to derive target functions for the intensities in crystallographic refinement, which in the acentric case results in a Rice distribution and in the centric case in a Woolfsson distribution (McCoy, 2004; Rupp, 2009; Murshudov et al., 2011).

In cryo-EM, a common choice for $T_{data}$ is to calculate the difference between the experimental map and the map calculated from the model in a least-squares sense, i.e., under the assumption that $\rho_{obs}(\mathbf{r})$ and $\rho_{calc}(\mathbf{r})$ are on the same scale, that

$$T_{data} = \int \left(\rho_{calc}(\mathbf{r}) - \rho_{obs}(\mathbf{r})\right)^2 \, \mathrm{d}\mathbf{r}, \tag{2.39}$$

where $\rho_{calc}(\mathbf{r})$ is the density calculated from the model and $\rho_{obs}(\mathbf{r})$ is the observed density from the cryo-EM reconstruction (Afonine et al., 2018b). Both $\rho_{calc}(\mathbf{r})$ and $\rho_{obs}(\mathbf{r})$ are typically represented on a Cartesian grid, with turns the integral in equation 2.39 into a sum over all grid points:

$$T_{data} = \sum_{i=1}^{N} \left(\rho_{calc}(\mathbf{r}_i) - \rho_{obs}(\mathbf{r}_i)\right)^2, \tag{2.40}$$

where $\rho_{calc}(\mathbf{r}_i)$ and $\rho_{obs}(\mathbf{r}_i)$ are the calculated and observed densities at grid point $i$, respectively. As the density in a real space experiment is not model biased like in reciprocal space, but rather directly observed (i.e. $\int \rho^2_{obs}(\mathbf{r})\,\mathrm{d}\mathbf{r} = \text{constant}$), as well as under the assumption that the overlap of the atomic densities does not change (i.e. $\int \rho^2_{calc}(\mathbf{r})\,\mathrm{d}\mathbf{r} = \text{constant}$), the target in equation 2.40 can be further simplified to

$$T_{data} = -\sum_{i=1}^{N} \rho_{calc}(\mathbf{r}_i)\rho_{obs}(\mathbf{r}_i). \tag{2.41}$$

For low-resolution structures, $\rho_{calc}(\mathbf{r})$ can be assumed to be almost constant over the volume of an atom, which means that $T_{data}$ in equation 2.41 can be further simplified to

$$T_{data} = -\sum_{i=1}^{N} \rho_{obs}(\mathbf{r}_i). \tag{2.42}$$

However, minimisation of equation 2.42 implies that the atoms in the model are pushed towards the highest local density (which might be shared by several atoms), which in turn implies the need for strong geometrical restraints in order to obtain a chemically sensible model from the refinement process (Rossmann et al., 2001; Afonine et al., 2018b).

An additional problem with cryo-EM structures is that the local resolution can, and does, vary significantly throughout the density, which means that some parts of the model are better supported by the data than others (Punjani et al., 2020). In order to account for the local resolution as well the thermal motion of the atoms in the model, a recent development for the calculation of $\rho_{atom}(\mathbf{r})$ is given by

$$\rho_{atom}(\mathbf{r}) = \sum_{i=1}^{M} C_i\,\Omega(\mathbf{r}, R_i, B_i), \tag{2.43}$$

where

$$\begin{aligned}\Omega(\mathbf{r}, R_i, B_i) =& \frac{1}{|\mathbf{r}|R_i}\left(\frac{1}{4\pi B_i}\right)^{1/2} \\ & \left(\exp\left(-\frac{4\pi^2(|\mathbf{r}| - R_i)^2}{B_i}\right) - \exp\left(\frac{4\pi^2(|\mathbf{r}| + R_i)^2}{B_i}\right)\right).\end{aligned} \tag{2.44}$$

The function $\Omega(\mathbf{r}, R_i, B_i)$ describes a spherically symmetric wave in real space, with a virtual unit charge that has been distributed uniformly over the surface of a sphere with radius $R_i$, blurred by an uncertainty parameter $B_i$ (Urzhumtsev and Lunin, 2022; Urzhumtsev et al., 2022; Urzhumtseva et al., 2023). The density, $\rho_{calc}(\mathbf{r})$, of a macromolecular model can then

be calculated as the sum of the atomic densities, $\rho_{atom}(\mathbf{r})$, over all atoms in the model, $N$, as

$$\rho_{calc}(\mathbf{r}) = \sum_{j=1}^{N} \rho_{atom_j}(\mathbf{r} - \mathbf{r}_j) = \sum_{j=1}^{N} \sum_{i=1}^{M} C_i \Omega(\mathbf{r} - \mathbf{r}_j, R_i, B_i), \tag{2.45}$$

The parameters $C_i$, $R_i$ and $B_i$ in equation 2.44 can be precomputed for all elements in the periodic table and stored in a lookup table for efficiency, allowing for fast calculation of $T_{data}$ in equation 2.40 during refinement (Urzhumtsev and Lunin, 2022; Urzhumtsev et al., 2022; Urzhumtseva et al., 2023).

### 2.3.1 The weight factor

The weight factor, $w_{data}$, in equation 2.31 balances the contribution of the data term, $T_{data}$, and the restraints term, $T_{restraints}$. If $w_{data}$ is too high, the model will overfit the data and may not be chemically sensible, while if $w_{data}$ is too low, the model will be overly constrained by the restraints and may not fit the data well. The question then becomes how to choose an appropriate value for $w_{data}$. A common suggestion is to choose $w_{data}$ such that the contributions from $T_{data}$ and $T_{restraints}$ to the total target function, $T$, are roughly equal based on their gradient norms, i.e.

$$w_{data} \approx \frac{||\nabla T_{restraints}||}{||\nabla T_{data}||}, \tag{2.46}$$

which has the effect of up-weighting the data term when the gradient of the restraints term is larger than the gradient of the data term and vice versa (Adams et al., 1997).

Ideally, the two terms on the right-hand side of 2.46 are congruent, at which point it could be argued that the prior is no longer needed or that the weight between them is arbitrary. In the case that they are not congruent, the effect of equation 2.46 is unfortunately that the term with a flatter gradient is up-weighted. Data of high resolution will for example have sharply defined peaks for the atomic positions, which implies that $||\nabla T_{data}||$ is large, which implies that $w_{data}$ is small and that the restraints term is up-weighted. The opposite is true for low-resolution data, where the peaks are broader and less defined, leading to a smaller $||\nabla T_{data}||$ and thus a larger $w_{data}$, which down-weights the restraints term. For this reason, the weight factor in equation 2.46 is often (at least in a crystallographic setting) prepended with a constant scale factor, usually set to $1/2$ (Jack and Levitt, 1978; Adams et al., 1997; McCoy, 2004; Fenn and Schnieders, 2011).

In order to determine $w_{data}$, a short MD simulation of the model is usually performed (Brünger et al., 1998; Liebschner et al., 2019). During this simulation, the fluctuations in

$T_{data}$ and $T_{restraints}$ are used to estimate $w_{data}$ as

$$w_{data} = \frac{\langle ||\nabla T_{restraints}|| \rangle}{\langle ||\nabla T_{data}|| \rangle}. \tag{2.47}$$

Other schemes to determine $w_{data}$ exist, which often relies on optimising some quality metric of the model, such as the $R_{free}$ factor in crystallography (see chapter 4, section 4.4) or cross-validation against a test set in cryo-EM (Afonine et al., 2011; Falkner and Schröder, 2013).

Interestingly, $w_{data}$ can be given physical meaning through statistical mechanics, as the probability of observing a particular atomic configuration can be related to the energy of that configuration, which is given by its Boltzmann weight, i.e.

$$P(\text{model}) \propto \exp\left(-\frac{E_{model}}{k_B T}\right), \tag{2.48}$$

where $E_{model}$ is the energy of the model, $k_B$ is the Boltzmann constant and $T$ is the temperature. By comparing equations 2.35 and 2.48, it is clear that the restraints term, $T_{restraints}$, can be interpreted as an energy term,

$$T_{restraints} = \frac{E_{model}}{k_B T}, \tag{2.49}$$

which in turn implies that the weight factor can be interpreted as a temperature-like parameter, i.e. $w_{data} = k_B T$ (Fenn and Schnieders, 2011).

# Chapter 3

# Computational chemistry

Chemistry is often thought of as an experimental science but it is also possible to study chemistry *in silico*, using computational methods to simulate and predict the chemistry of a system. In this chapter I will give an overview of the computational methods used in this thesis, more specifically quantum mechanics (QM), molecular mechanics (MM) as well as the combination of methods with different levels of complexity.

## 3.1 Quantum chemistry

Quantum mechanics is a fundamental theory of matter and energy on the atomic and subatomic level. It is based on the idea that energy, momentum and other quantities are quantised and can only take on discrete values. At the core of quantum mechanics is the wave function which describes the state of a system and in turn the wave function contains all the observable information about the system through the use of operators.

### 3.1.1 The Schrödinger equation

The notion of a wave function, together with the now famous Schrödinger equation, was introduced by Erwin Schrödinger in 1926 in response to the postulate made by Louis de Broglie in 1924 that matter has wave-like properties (de Broglie, 1925; Schrödinger, 1926b). In its most general form, the time-dependent Schrödinger equation for a particle moving in three dimensions is given by

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = \hat{\mathrm{H}}\Psi(\mathbf{r},t), \tag{3.1}$$

where $\hbar$ is the reduced Planck constant, $\Psi(\mathbf{r}, t)$[1] is the wave function of the system and $\hat{\mathrm{H}}$ is the Hamiltonian operator which describes the total energy, $E$, of the system, for a point $\mathbf{r}$ at time $t$. Through separation of variables, the time-independent form of the Schrödinger equation for the stationary states of the system can be obtained as

$$\hat{\mathrm{H}}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}), \tag{3.2}$$

which is an eigenvalue equation where the wave function $\Psi(\mathbf{r})$ is the eigenfunction in the form of a standing wave and the energy $E$ is the corresponding eigenvalue (Schrödinger, 1926a). According to the Born interpretation of quantum mechanics, the probability of finding a particle at position $\mathbf{r}$ is proportional to $|\Psi(\mathbf{r})|^2$ (Born, 1926).

As in classical mechanics, the quantum mechanical Hamiltonian operator can be decomposed into separate components, representing the kinetic and potential energies of the particles in the system:

$$\hat{\mathrm{H}} = \hat{\mathrm{T}}_e + \hat{\mathrm{T}}_n + \hat{\mathrm{V}}_{en} + \hat{\mathrm{V}}_{ee} + \hat{\mathrm{V}}_{nn}, \tag{3.3}$$

where $\hat{\mathrm{T}}_e$ is the kinetic energy of the electrons, $\hat{\mathrm{T}}_n$ is the kinetic energy of the nuclei, $\hat{\mathrm{V}}_{en}$ is the electron–nuclear attraction, $\hat{\mathrm{V}}_{ee}$ is the electron–electron repulsion, $\hat{\mathrm{V}}_{nn}$ is the nuclear–nuclear repulsion. For a set of $N$ electrons and $M$ nuclei, the different components of the Hamiltonian operator in equation 3.3 are given by

$$\hat{\mathrm{T}}_e = -\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2 \tag{3.4}$$

and

$$\hat{\mathrm{T}}_n = -\sum_{I=1}^{M}\frac{\hbar^2}{2m_I}\nabla_I^2, \tag{3.5}$$

for the kinetic energies, where $m_e$ is the mass of an electron, $m_I$ is the mass of nucleus $I$, $\nabla_i^2$ and $\nabla_I^2$ are the Laplace operators with respect to the coordinates of electron $i$ and nucleus $I$, respectively, and

$$\hat{\mathrm{V}}_{en} = -\frac{e^2}{4\pi\varepsilon_0}\sum_{i=1}^{N}\sum_{I=1}^{M}\frac{Z_I}{|\mathbf{r}_i - \mathbf{r}_I|}, \tag{3.6}$$

$$\hat{\mathrm{V}}_{ee} = \frac{e^2}{4\pi\varepsilon_0}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{3.7}$$

[1]In this chapter, as opposed to reciprocal space in chapter 2.1, the use of the notation $\mathbf{r}$ and $\mathbf{r}_n$ refers to three-dimensional Cartesian coordinates, not necessarily fractional ones.

and

$$\hat{\mathrm{V}}_{nn} = \frac{e^2}{4\pi\varepsilon_0} \sum_{I=1}^{M} \sum_{J>I}^{M} \frac{Z_I Z_J}{|\mathbf{r}_I - \mathbf{r}_J|} \tag{3.8}$$

for the potentials, where $e$ is the elementary charge, $\varepsilon_0$ is the permittivity of free space and $Z_I$ is the atomic number of nucleus $I$.

Under the Born–Oppenheimer approximation, where the nuclei are assumed to be stationary owing to their much larger mass compared to the electrons, the kinetic energy of the nuclei can be neglected, i.e. $\hat{\mathrm{T}}_n = 0$ (Born and Oppenheimer, 1927). Similarly, the nuclear–nuclear repulsion term, $\hat{\mathrm{V}}_{nn}$, is constant for a given arrangement of nuclei and can thus be treated as a shift in energy given by $\langle\Psi|\hat{\mathrm{V}}_{nn}|\Psi\rangle = V_{nn}$. If needed, it can be added back at the end of a calculation. The Hamiltonian operator for an electronic structure calculation, which is the main interest in quantum chemistry, is under these assumptions given by

$$\hat{\mathrm{H}}_{elec} = \hat{\mathrm{T}}_e + \hat{\mathrm{V}}_{en} + \hat{\mathrm{V}}_{ee}, \tag{3.9}$$

which while reduced in complexity compared to the full Hamiltonian in equation 3.3 still results in a very complicated partial differential equation that cannot be solved analytically for systems with more than one electron.

### 3.1.2 Hartree–Fock theory

One of the simplest methods to approximately solve the electronic Schrödinger equation for a many-body system is the Hartree–Fock (HF) method (Hartree, 1928a; Hartree and Hartree, 1935; Slater, 1951). With a general coordinate $\mathbf{x}_i = (\mathbf{r}_i, \omega_i)$, which includes both spatial, $\mathbf{r}_i$, and spin coordinates, $\omega_i$, for electron $i$, the HF method starts with the ansatz that the many-electron wave function for the entire system can be approximated by a product of orthonormal one-electron wave functions, $\phi_i(\mathbf{x}_i)$, usually called "spin-orbitals", i.e. that

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \approx \phi_1(\mathbf{x}_1)\phi_2(\mathbf{x}_2)\ldots\phi_N(\mathbf{x}_N), \tag{3.10}$$

which is known as a "Hartree product". The spin-orbitals are in turn given by

$$\phi_i(\mathbf{x}_i) = \psi_i(\mathbf{r}_i)\sigma(\omega_i), \tag{3.11}$$

where $\psi_i(\mathbf{r}_i)$ is a spatial orbital and $\sigma(\omega_i)$ is a spin function, which can take on one of two values for electrons, typically denoted $\alpha(\omega_i)$ or $\beta(\omega_i)$ (where $\langle\alpha|\alpha\rangle = \langle\beta|\beta\rangle = 1$ and $\langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle = 0$), corresponding to spin-up ($m_s = 1/2$) or spin-down ($m_s = -1/2$), respectively.

However, this ansatz does not account for the Pauli principle, which states that no two

fermions (such as electrons) can occupy the same quantum state simultaneously (Slater, 1930a; Fock, 1930). The wave function must thus be antisymmetric with respect to the exchange of any two electrons, i.e.

$$\Psi(\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_j,\ldots) = -\Psi(\ldots,\mathbf{x}_j,\ldots,\mathbf{x}_i,\ldots). \tag{3.12}$$

Taking a linear combination of all possible permutations of the Hartree product in equation 3.10 with alternating signs results in a wave function that satisfies the antisymmetry requirement in equation 3.12 and is known as a Slater determinant (SD):

$$\Psi(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)^{SD} = \frac{1}{\sqrt{N!}}\begin{vmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \ldots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \ldots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \ldots & \phi_N(\mathbf{x}_N) \end{vmatrix}; \langle\phi_i|\phi_j\rangle = \delta_{ij}, \tag{3.13}$$

where $1/\sqrt{N!}$ is a normalisation constant (Heisenberg, 1926; Slater, 1929; Dirac, 1926). In HF theory, the trial wave function is taken to be a single SD which also implies that the electrons are treated as independent particles moving in the average field created by all other electrons, or in other words, under this assumption electron correlation is neglected and thus the HF method is a mean-field approximation.

The spin-orbitals constituting the trial wave function constructed from a SD in equation 3.13 can in turn be determined through use of the variational principle, which states that the expectation value of the energy for any trial wave function is always greater than or equal of the true ground-state energy of the system, i.e.

$$E_{trial} = \frac{\langle\Psi_{trial}|\hat{\mathrm{H}}|\Psi_{trial}\rangle}{\langle\Psi_{trial}|\Psi_{trial}\rangle} \geq E_0, \tag{3.14}$$

where $E_0$ is the true ground-state energy of the system. Application of the variational principle to a SD results in the HF equations (Slater, 1928; Gaunt, 1928).

In order to simplify the notation, Dirac's bra-ket notation will be used in the following (Dirac, 1939). Additionally, $j_0 = e^2/(4\pi\varepsilon_0)$ will be used to simplify the equations.

The HF equations in their canoncical form are given by

$$\hat{\mathrm{f}}|\phi_i\rangle = \varepsilon_i|\phi_i\rangle, \tag{3.15}$$

where $\hat{\mathrm{f}}$ is the Fock operator and $\varepsilon_i$ is the orbital energy of spin-orbital $i$. The Fock operator is given by

$$\hat{\mathrm{f}} = \hat{\mathrm{h}} + \sum_{j=1}^{N}\left(\hat{\mathrm{J}}_j - \hat{\mathrm{K}}_j\right), \tag{3.16}$$

where

$$\hat{\mathrm{h}} = -\frac{\hbar^2}{2m_e}\nabla^2 - j_0 \sum_{I=1}^{M} \frac{Z_I}{|\mathbf{r} - \mathbf{r}_I|} \tag{3.17}$$

is the one-electron operator (or "core Hamiltonian"), which includes the kinetic energy of the electron and its attraction to the nuclei,

$$\hat{\mathrm{J}}_j|\phi_i\rangle := j_0 \left( \int \frac{\phi_j^*(\mathbf{x}')\phi_j(\mathbf{x}')}{|\mathbf{r} - \mathbf{r}'|}\, \mathrm{d}\mathbf{x}' \right) |\phi_i\rangle = j_0 \langle \phi_j \left| |r|^{-1} \right| \phi_j \rangle |\phi_i\rangle \tag{3.18}$$

defines the Coulomb operator, which describes the average repulsion between the electron in spin-orbital $i$ and the electron in spin-orbital $j$, and

$$\hat{\mathrm{K}}_j|\phi_i\rangle := j_0 \left( \int \frac{\phi_j^*(\mathbf{x}')\phi_i(\mathbf{x}')}{|\mathbf{r} - \mathbf{r}'|}\, \mathrm{d}\mathbf{x}' \right) |\phi_j\rangle = j_0 \langle \phi_j \left| |r|^{-1} \right| \phi_i \rangle |\phi_j\rangle \tag{3.19}$$

defines the exchange operator, which describes the exchange interaction between electrons due to the antisymmetry requirement of the wave function. It does not have any classical analogue.

The total (HF) energy of the system is given by

$$E^{HF} = \sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \sum_{j>i}^{N} \left( J_{ij} - K_{ij} \right), \tag{3.20}$$

where

$$J_{ij} = \langle \phi_i(\mathbf{x}) | \hat{\mathrm{J}}_j | \phi_i(\mathbf{x}) \rangle = j_0 \langle \phi_i(\mathbf{x}) \phi_j(\mathbf{x}') \left| |r|^{-1} \right| \phi_i(\mathbf{x}) \phi_j(\mathbf{x}') \rangle \tag{3.21}$$

$$K_{ij} = \langle \phi_i(\mathbf{x}) | \hat{\mathrm{K}}_j | \phi_i(\mathbf{x}) \rangle = j_0 \langle \phi_i(\mathbf{x}) \phi_j(\mathbf{x}') \left| |r|^{-1} \right| \phi_j(\mathbf{x}) \phi_i(\mathbf{x}') \rangle \tag{3.22}$$

are the Coulomb and exchange integrals, respectively, and

$$\varepsilon_i = \langle \phi_i | \hat{\mathrm{f}} | \phi_i \rangle = \langle \phi_i | \hat{\mathrm{h}} | \phi_i \rangle + \sum_{j=1}^{N} (J_{ij} - K_{ij}). \tag{3.23}$$

While the HF equations are similar in appearance to standard eigenvalue problems, these equations are now a set of $N$ coupled nonlinear integro-differential equations, as the Fock operator, $\hat{\mathrm{f}}$, for each spin-orbital depends on all other $N - 1$ spin-orbitals. Thus, the HF equations must be solved iteratively in a self-consistent field (SCF) manner, where an initial guess for the spin-orbitals is made, the Fock operators are constructed, the HF equations are solved to obtain a new set of spin-orbitals, which are then used as input for the next iteration and this is then repeated until some convergence criterion is met (Hartree, 1928b).

The HF equations can be solved numerically by mapping the orbitals onto a grid in real space (Kobus, 2013). However, the more common approach is to use a basis set expansion. In the basis set expansion approach, the spatial part of the spin-orbitals are expressed as linear combinations of a finite set of $M_{basis}$ known basis functions, $\theta_j(\mathbf{r})$, i.e.

$$\psi_i(\mathbf{r}) = \sum_{j=1}^{M_{basis}} c_{ji}\theta_j(\mathbf{r}), \tag{3.24}$$

where $c_{ji}$ are the expansion coefficients. An infinite number of basis functions would result in the HF limit, i.e. the variationally best single-determinant wave function that can be achieved, but is for obvious reasons computationally intractable (Jensen, 2017).

The introduction of a basis set also allows the HF equations to be cast into a matrix form, resulting in the Roothaan–Hall equations:

$$\mathbf{FC} = \mathbf{SC}\varepsilon, \tag{3.25}$$

where $\mathbf{F}$ is the Fock matrix ($F_{ij} = \langle\phi_i|\hat{\mathrm{f}}|\phi_j\rangle$), $\mathbf{C}_i$ is the coefficient vector for the $i$-th spatial orbital, $\varepsilon_{ii}$ contains the corresponding orbital energy and $\mathbf{S}$ is the overlap matrix ($S_{ij} = \langle\theta_i|\theta_j\rangle$) (Roothaan, 1951; Hall, 1951; Jensen, 2017). In turn, this transforms the integro-differential HF equations into a pseudo-matrix eigenvalue problem, which like the HF equations can be solved iteratively in an SCF manner (Jensen, 2017).

Formally, the time complexity of HF scales as $\mathcal{O}(N^4)$ with the number of basis functions used, owing to the need to calculate the two-electron integrals for the Coulomb and exchange operators in equations 3.21 and 3.22.

### 3.1.3 Basis sets

The set of basis functions in equation 3.24 can in principle be any set of functions that, ideally, are complete and orthonormal. A specific set of basis functions is within the field of computational chemistry often referred to as a "basis set". However, in practise the basis functions are typically chosen to resemble the atomic orbitals of the atoms in the system, which is known as the Linear Combination of Atomic Orbitals (LCAO) approach, in order to make the calculations needed tractable (Jensen, 2017). In the LCAO approach, the basis functions are centered on the nuclei of the atoms in the system and can be classified into two main types: Slater-type orbitals (STOs) and Gaussian-type orbitals (GTOs) (Slater, 1930b; Boys, 1950).

For STOs, the radial part of a basis function is given by

$$R_n(r) = Nr^{n-1}e^{-\zeta r}. \tag{3.26}$$

For GTOs, the radial part of a basis function is given by

$$R_n(r) = Nr^{n-1}e^{-\zeta r^2}. \tag{3.27}$$

In both equations, $N$ represents a normalisation constant, $n$ is the principal quantum number, $r$ is the distance from the nucleus and $\zeta$ is a parameter that controls the width of the orbital. Both types of basis functions can also include angular parts given by spherical harmonics, $Y_{lm}(\theta, \phi)$, where $l$ is the azimuthal quantum number, $m$ is the magnetic quantum number and $0 \leq \theta \leq \pi$ and $0 \leq \phi < 2\pi$ are the polar and azimuthal angles, respectively.

STOs have the correct asymptotic behaviour both at the nucleus (the Kato cusp condition) and at large distances from the nucleus, but the integrals required for HF calculations with STOs are computationally expensive. GTOs on the other hand do not display the correct behaviour at the nucleus and they also decay too rapidly, making them a worse approximation of the true atomic orbitals. However, the integrals required for HF calculations with GTOs can be evaluated efficiently and owing to the fact that the product of two Gaussian functions is another Gaussian function, contracted Gaussians can be formed by taking linear combinations of multiple GTOs to better approximate the shape of STOs (Atkins and Friedman, 2010). The introduction of GTOs in 1950 by S. F. Boys is generally hailed as a major breakthrough in computational chemistry as it allowed for efficient evaluation of the integrals needed for HF calculations (Boys, 1950; Kamberaj, 2023).

Basis sets can also be classified based on the number of basis functions used to describe each atomic orbital. A minimal basis set uses one basis function for each atomic orbital, while a double-zeta basis set uses two basis functions for each atomic orbital, a triple-zeta basis set uses three, and so on. Additionally, polarisation functions, which are basis functions of higher angular momentum than those occupied in the ground state (e.g. *p*-functions for *s*-electrons, *d*-functions for *p*-electrons, etc.), can be added to the basis set to allow for more flexibility in the shape of the orbitals (Pitman et al., 2023). To better describe the electron density in regions far from the nucleus, diffuse functions, which are Gaussian basis functions with small exponents (i.e. they are more spread out in space), can be added to the basis set as well (Pitman et al., 2023).

In this thesis, the def2-SV(P) basis set from the Karlsruhe family has been used exclusively, which is a split-valence basis set (i.e. the core orbitals are represented with one basis function whereas the valence orbitals are represented with two basis functions), with polarisation functions added for non-hydrogen atoms (Weigend and Ahlrichs, 2005).

### 3.1.4 Density functional theory

Another significant breakthrough for computational chemistry came in 1964 when Hohenberg and Kohn proved two theorems that would form the basis of density functional theory (DFT) (Hohenberg and Kohn, 1964; Fransson et al., 2022). The Hohenberg–Kohn theorems state that:

- **First theorem:** The external potential, $V_{ext}(\mathbf{r})$, of a given system is a unique functional (up to a trivial additive constant) of the electron density, $\rho(\mathbf{r})$. Since the Hamiltonian is determined by the external potential, it follows that all properties of the system thus are determined by the electron density alone.
  **Corollary:** The ground-state electron density uniquely determines the ground-state wave function and therefore also all observables of the system.
- **Second theorem:** The ground-state energy can be obtained variationally, i.e. the electron density that minimises the energy is the true ground-state electron density.
  **Corollary:** For any trial density $\tilde{\rho}(\mathbf{r})$ that satisfies $\tilde{\rho}(\mathbf{r}) \geq 0$ and $\int \tilde{\rho}(\mathbf{r})\,\mathrm{d}\mathbf{r} = N$, where $N$ is the number of electrons, the energy functional satisfies $E[\tilde{\rho}] \geq E[\rho]$.

Interestingly, this also implies that while the (electronic) wave function is a function of $3N$ spatial coordinates and $N$ spin coordinates, where $N$ is the number of electrons, the electron density is a function of only 3 spatial coordinates, but still contains the same information as the wave function.

The ground-state electronic energy of the system can in turn be expressed exactly as a functional of the electron density:

$$E[\rho] = T[\rho] + V_{ne}[\rho] + V_{ee}[\rho] = \int \rho(\mathbf{r}) V_{ext}(\mathbf{r})\,\mathrm{d}\mathbf{r} + F[\rho], \tag{3.28}$$

where $T[\rho]$ is the kinetic energy functional, $V_{ne}[\rho]$ is the electron–nuclear attraction functional, $V_{ee}[\rho]$ is the electron–electron repulsion functional and $F[\rho] = T[\rho] + V_{ee}[\rho]$ is called the "universal functional" and does not depend on the external potential. The external potential is typically given by the electron–nuclear attraction potential for a given arrangement of nuclei (in the absence of any external fields), i.e. $V_{ext}(\mathbf{r}) = V_{en}(\mathbf{r})$.

Opting for this route to obtain the ground-state energy seems quite appealing, as the problem is reduced to finding the electron density that minimises the energy functional in equation 3.28. However, while the Hohenberg–Kohn theorems state that the electron density functional exists, the theorems unfortunately do not provide an analytical form for the functional and attempts at this orbital-free version of DFT generally performs poorly, mainly due to the approximations made for the kinetic energy functional (García-Aldea

and Alvarellos, 2007).

In 1965, Kohn and Sham proposed a method to circumvent this problem by re-introducing the use of orbitals through the use of auxiliary non-interacting one-electron functions, that by definition has the same electron density as the real, interacting, system, i.e.

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2, \tag{3.29}$$

where $\psi_i(\mathbf{r})$ are the one-electron spatial orbitals of the non-interacting system, similar to HF (Kohn and Sham, 1965). The kinetic energy functional for these orbitals can be calculated exactly and is given by

$$T_s[\rho] = -\frac{\hbar^2}{2m_e} \sum_{i=1}^{N} \langle \psi_i | \nabla^2 | \psi_i \rangle, \tag{3.30}$$

which to within a small residual correction term is a good approximation to the true kinetic energy functional, i.e. $T[\rho] = T_s[\rho] + T_c[\rho] \approx T_s[\rho]$. Kohn and Sham then defined that

$$F[\rho] := T_s[\rho] + J[\rho] + E_{xc}[\rho], \tag{3.31}$$

where $J[\rho]$ is the classical Coulomb energy, given by

$$J[\rho] = \frac{1}{2} \frac{e^2}{4\pi\varepsilon_0} \int \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}\mathbf{r} \, \mathrm{d}\mathbf{r}', \tag{3.32}$$

and $E_{xc}[\rho]$ is the exchange–correlation energy functional, which captures all non–classical effects of the electron–electron interaction as well as the kinetic energy correction (Fransson et al., 2022).

Given that $F[\rho] = T[\rho] + V_{ee}[\rho]$ it follows that

$$E_{xc}[\rho] = (T[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{3.33}$$

which unfortunately is an unknown functional of the electron density. Several approaches exist to approximate this functional (Kohn et al., 1996; Perdew et al., 1996, 2009). In the local density approximation (LDA), $E_{xc}[\rho]$ is assumed to be a functional of the local electron density only, i.e.

$$E_{xc}^{LDA}[\rho] = \int \rho(\mathbf{r}) \epsilon_{xc}(\rho(\mathbf{r})) \, \mathrm{d}\mathbf{r}. \tag{3.34}$$

In the generalised gradient approximation (GGA), $E_{xc}[\rho]$ is assumed to be a functional of both the local electron density and its gradient, i.e.

$$E_{xc}^{GGA}[\rho] = \int \rho(\mathbf{r}) \epsilon_{xc}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})) \, \mathrm{d}\mathbf{r}. \tag{3.35}$$

Attempts at more advanced exchange–correlation functionals also exist, such as in the meta-GGA approach, where the Laplacian of the electron density and/or the kinetic energy density, $\tau$, are included as additional variables, i.e.

$$E_{xc}^{mGGA}[\rho] = \int \rho(\mathbf{r})\epsilon_{xc}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}), \nabla^2\rho(\mathbf{r}), \tau(\mathbf{r}))\,\mathrm{d}\mathbf{r}. \tag{3.36}$$

A fourth class are hybrid functionals, which mix in a portion of exact exchange from HF with DFT exchange–correlation functionals, i.e.

$$E_{xc}^{hybrid}[\rho] = E_{xc}^{DFT}[\rho] + \gamma E_{x}^{HF}[\rho], \tag{3.37}$$

of which the famous B3LYP functional is an example (Vosko et al., 1980; Lee et al., 1988; Becke, 1988; Stephens et al., 1994).

The exact, and total, ground-state electronic energy functional in the Kohn–Sham (KS) approach is then given by

$$E[\rho] = \int \rho(\mathbf{r})V_{ext}(\mathbf{r})\mathrm{d}\mathbf{r} + T_s[\rho] + J[\rho] + E_{xc}[\rho], \tag{3.38}$$

which through the introduction of orbitals in equation 3.29 transforms the problem of finding the electron density that minimises the energy functional into finding the set of coefficients for a given basis set that minimises the energy functional through the Kohn–Sham equations, which in their canonical form are given by

$$\left(-\frac{\hbar^2}{2m_e}\nabla^2 + V_{eff}\right)\psi_i(\mathbf{r}) = \varepsilon_i\,\psi_i(\mathbf{r}), \tag{3.39}$$

where $V_{eff}$ is the effective potential, given by

$$V_{eff} = \frac{e^2}{4\pi\varepsilon_0}\int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}' + V_{xc}(\mathbf{r}) + V_{ext}(\mathbf{r}), \tag{3.40}$$

where $V_{xc}(\mathbf{r}) = \delta E_{xc}[\rho]/\delta\rho(\mathbf{r})$ is the exchange–correlation potential and the only unknown term in the above equation.

Similarly to HF, equations 3.39 can be solved self-consistently in order to obtain the expansion coefficients for the chosen basis set, with a time complexity that scales as $\mathcal{O}(N^3)$, where $N$ is the number of basis functions used, which allows for calculations on larger systems compared to HF.

With care taken when chosing the basis set and functional, DFT typically yield excellent geometries and reasonable energies (Neese, 2006; Benediktsson and Bjornsson, 2022; Vysotskiy et al., 2023). In this thesis, the meta-GGA functional TPSS has been used exclusively, together with the empirical DFT–D4 dispersion correction, as both HF and DFT have in common that they do not properly treat dispersion interactions (Tao et al., 2003; Caldeweyher et al., 2019).

## 3.2 Molecular mechanics

In molecular mechanics, the atoms in a system are treated as classical particles, i.e. as point masses, and the interactions between them are described by empirical potential functions. The total potential energy of the system is then a function of the positions of all atoms in the system, i.e. $E_{total} = E_{total}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$.

The interactions in the system are commonly divided into bonded and non-bonded interactions, i.e.

$$E_{total} = E_{bonded} + E_{non-bonded}. \tag{3.41}$$

The bonded terms are commonly assumed to consist of covalent bonds, bond angles and dihedral angles, where the bonds and angles are usually modelled as harmonic oscillators and the dihedral angles as periodic functions,

$$\begin{aligned} E_{bonded} &= E_{bonds} + E_{angles} + E_{dihedrals} \\ &= \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} V_n \left(1 + \cos(n\phi - \gamma)\right), \end{aligned} \tag{3.42}$$

where $k_b$ is the bond force constant, $r$ is the current bond length, $r_0$ is the equilibrium bond length, $k_\theta$ is the angle force constant, $\theta$ is the current bond angle, $\theta_0$ is the equilibrium bond angle, $V_n$ is the barrier height, $n$ is the periodicity, $\phi$ is the current dihedral angle and $\gamma$ is the phase shift.

The non-bonded terms typically consist of a Lennard–Jones potential for the van der Waals interactions and a Coulomb potential for the electrostatic interactions,

$$\begin{aligned} E_{non-bonded} &= E_{vdW} + E_{electrostatic} \\ &= \sum_{i=1}^{N} \sum_{j>i}^{N} 4\varepsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{4\pi\varepsilon_r \varepsilon_0 r_{ij}}, \end{aligned} \tag{3.43}$$

where $\varepsilon_{ij}$ is the depth of the potential well, $\sigma_{ij}$ is the distance at which the inter-particle potential is zero, $r_{ij}$ is the current distance between atoms $i$ and $j$, $q_i$ and $q_j$ are the partial charges of atoms $i$ and $j$, respectively, $\varepsilon_r$ is the relative permittivity of the medium and $\varepsilon_0$ is the permittivity of free space. Calculation of the non-bonded terms is the limiting factor for MM calculations and scales as $\mathcal{O}(N^2)$ in a naïve implementation, where $N$ is the number of atoms in the system, which can be computationally expensive for large systems. To reduce the computational cost, a cutoff distances can be introduced, beyond which the non-bonded interactions are neglected. In neutral and periodic systems, Ewald summation can be used to accelerate the calculations (Ewald, 1921).

While all the constants in equations 3.42 and 3.43 are in theory unique for each pair (bonds), triplet (angles) or quadruplet (dihedrals) of atoms, in practise they are grouped into types based on the chemical environment of the atoms involved, in order to reduce the number of parameters needed. The values of the constants for the types are derived from experimental data or high-level quantum mechanical calculations (Wang et al., 2004, 2006). A specific set of constants in equations 3.42 and 3.43 is usually referred to as a "force field". A disadvantage with MM methods is that they in general cannot describe bond breaking or formation, as the bonds are treated as fixed entities. While reactive force fields do exist, they are not widely used (Brenner et al., 2002; Senftle et al., 2016; Leven et al., 2021).

In macromolecular structure refinement, force fields are commonly used as a way to encode prior chemical knowledge about the system in the form of geometric restraints, in order to guide the refinement towards chemically reasonable structures and help mitigate the low data-to-parameter ratio characteristic of macromolecular structure determination (Levitt and Lifson, 1969; Jack and Levitt, 1978; Brünger et al., 1998; Engh and Huber, 1991; Adams et al., 2002; Moriarty et al., 2020). From comparing equations 3.42 and 2.37, we see that for example $k_b \propto 1/\sigma_b^2$ and $r_0 = r_{target}$, i.e. the force constant is inversely proportional to the variance of the restraint and the equilibrium bond length is equal to the target value of the restraint.

## 3.3 Hybrid methods

In general, high accuracy from a computational method comes with the tradeoff of increased computational cost. Typically, at least in computational chemistry, not all parts of a system are equally important, and it is thus of interest to treat the interesting parts with a high-accuracy method (e.g. QM), while the less important regions can then be treated with a lower-accuracy method (e.g. MM), in order to save computational resources or to even make the calculation tractable.

Such a scheme was suggested in 1976 by Warshel and Levitt, where QM was employed for the active site of an enzyme while the rest of the protein and the surrounding solvent was treated with MM (Warshel and Levitt, 1976). Combination of these two methods is usually referred to as "QM/MM", but can also be extended to other combinations of methods with different levels of complexity. Hybrid schemes are in particular of interest for proteins containing metals, as the MM parametrisations are often poor or non-existing for metals, while QM methods can accurately describe the chemistry in these cases (Mackerell, 2004; Senn and Thiel, 2009).

In the following, "*high*" refers to any high-accuracy method, while "*low*" refers to any low-accuracy method. The use of "*real*" refers to the whole model, whereas "*model*" refers to a region of interest (*model* is thus a subset of *real*). A capital letter in bold face refers to a set of coordinates, where the subscript indicates to which region the coordinates belong, e.g. $\mathbf{R}_{real}$ refers to the coordinates of the entire system, while $\mathbf{R}_{model}$ refers to the coordinates of the *model* region.

Generally, two main schemes for hybrid methods exist, the additive and the subtractive scheme. In the additive scheme, the total energy of the system is given by

$$\begin{aligned} E_{total}(\mathbf{R}_{real}) =& E_{high,model}(\mathbf{R}_{model}) + E_{low,real-model}(\mathbf{R}_{real-model}) \\ &+ E_{int(model,real)}(\mathbf{R}_{model}, \mathbf{R}_{real}), \end{aligned} \tag{3.44}$$

where $E_{int(model,real)}$ is an interaction term between the two regions. The additive scheme requires specialised software, as the interaction term must be calculated in a way that is consistent with both methods, as well as it must be possible to exactly pick which energy terms to include in the lower accuracy region (as essentially a "hole" is created in the *low*-accuracy region where the *high*-accuracy region is located) (Cao and Ryde, 2018).

In the subtractive scheme, the total energy of the system is given by

$$E_{total}(\mathbf{R}_{real}) = E_{low,real}(\mathbf{R}_{real}) + E_{high,model}(\mathbf{R}_{model}) - E_{low,model}(\mathbf{R}_{model}), \tag{3.45}$$

where the subtraction is needed to avoid double counting the energy of the model region.

The subtractive scheme has the advantage that it can be implemented in any MM software that can calculate the energy of the whole system, as well as the energy of a smaller region, and no special interaction term is needed. However, the subtractive scheme requires that a parametrisation for the model region exists, which is not always the case (e.g. for unusual ligands or metal ions), although dummy (i.e. zeroed) parameters can be used in such cases. Proper care needs to be taken when cutting covalent bonds between the model and the real region as well, as improper handling of this can introduce artefacts. Implemented properly, the additive and subtractive schemes give very similar results (Cao and Ryde, 2018).

Both additive and subtractive schemes can be further extended to multiple layers, where different regions are treated with different methods of varying accuracy. An $n$-layer subtractive scheme would for example be given by

$$E_{total}(\mathbf{R}_{real}) = E_{low,real}(\mathbf{R}_{real}) + \sum_{i=1}^{n-1} (E_{i,model_i}(\mathbf{R}_{model_i}) - E_{i+1,model_i}(\mathbf{R}_{model_i})), \tag{3.46}$$

where $model_i$ is the region treated with method $i$, with method 1 being the highest accuracy method and method $n$ being the lowest accuracy method (Chung et al., 2015).

The treatment of the boundary between different regions is an important aspect of hybrid methods. In the case of QM/MM, there are typically three schemes to treat the electrostatic interaction between the QM and MM regions (Lin and Truhlar, 2007):

- **Mechanical embedding**, where the electrostatic interactions between the QM and MM regions are treated at the MM level only, through point charges, also for the model region. This is the simplest scheme and also the least accurate as no polarisation between the QM and MM regions is accounted for.
- **Electrostatic embedding**, where MM point charges outside of the QM region are included in the QM Hamiltonian, thus allowing for polarisation of the QM region by the MM region. This is a more accurate scheme, but also more computationally demanding as the QM Hamiltonian must be modified to include the MM point charges.
- **Polarisable embedding**, extends the electrostatic embedding scheme and accounts for polarisation of the MM region by the QM region. This is the most accurate scheme but also the most complex as it requires a polarisable force field. Additionally, the polarisable embedding scheme is more complex to implement as it requires a self-consistent treatment of the polarisation between the QM and MM regions in the QM calculations.

When cutting covalent bonds, two main approaches exist:

- **Link atoms**, where the cut bond is capped with a link atom, typically a hydrogen atom, in order to saturate the valence of the cut bond. The position of the link atom is usually determined based on the positions of the two atoms forming the original bond (Field et al., 1990; Ryde, 1996; Ryde and Olsson, 2001).
- **Localised orbitals**, where localised orbitals that represent the electronic structure at the boundary are used and kept frozen during the SCF calculation. This approach avoids the introduction of artificial link atoms, but requires more complex handling of the boundary region (Murphy et al., 2000).

A subtractive hybrid scheme, with no electrostatic interaction between the high and low regions, utilising hydrogen link atoms, has been used in this thesis.

### 3.3.1 The link atom approach

In the link atom approach, a link atom, $L_{model}$, is positioned along the bond vector of each cut bond, where the position of the link atom, $\mathbf{r}_L$, is given by

$$\mathbf{r}_L = \mathbf{r}_{X_{model}} + g_{bond}(\mathbf{r}_{Y_{real}} - \mathbf{r}_{X_{model}}), \tag{3.47}$$

where $\mathbf{r}_{X_{real}} = \mathbf{r}_{X_{model}}$ and $\mathbf{r}_{Y_{real}}$ are the positions of the two atoms, $X_{real} = X_{model}$ and $Y_{real}$, forming the original bond being cut, see figure 3.1.

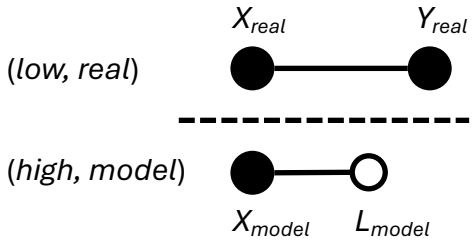



Figure 3.1: Schematic illustration of the link atom approach for cutting a covalent bond between the *model* and *real* regions. The link atom, $L_{model}$, is positioned along the bond vector between atoms $X_{real}$ and $Y_{real}$, scaled by the factor $g_{bond}$.

The scaling factor, $g_{bond}$, which will be unique for each type of cut bond, is determined as

$$g_{bond} = \frac{|\mathbf{r}_L - \mathbf{r}_{X_{model}}|}{|\mathbf{r}_{Y_{real}} - \mathbf{r}_{X_{model}}|} = \frac{d_{X-L}}{d_{X-Y}}, \tag{3.48}$$

where $d_{X-L}$ is the distance from the *model* atom to the link atom as given by the higher level of theory and $d_{X-Y}$ is the distance from the *model* atom to the *real* atom as given by the lower level of theory. The distance $d_{X-L}$ is typically precalculated from a QM optimisation and stored in database or lookup table, whereas $d_{X-Y}$ is obtained directly from the force field parametrisation in the form of an ideal bond length.

Through use of the chain rule, the gradient of the total energy in the subtractive scheme in equation 3.45 is then given by

$$\begin{aligned}\nabla E_{total}(\mathbf{R}_{real}) =& \nabla E_{low,real}(\mathbf{R}_{real}) + \nabla E_{high,model}(\mathbf{R}_{model})J(\mathbf{R}_{model}\,;\mathbf{R}_{real}) \\ &- \nabla E_{low,model}(\mathbf{R}_{model})J(\mathbf{R}_{model}\,;\mathbf{R}_{real}),\end{aligned} \tag{3.49}$$

where $J(\mathbf{R}_{model}\,;\mathbf{R}_{real})$ is the Jacobian matrix between $\mathbf{R}_{model}$ and $\mathbf{R}_{real}$, which for the link atoms is calculated by taking the appropriate derivatives of equation 3.47. For all atoms inside the *model* region, the corresponding block in the Jacobian matrix will be the identity

matrix, as their positions are the same in both regions. Likewise, for all atoms in the *real* region outside of the *model* region, the corresponding block in the Jacobian matrix will be the zero matrix. For each cut bond, $(1 - g_{bond})$ of the gradient of the link atom will be projected onto $X_{real}$, whereas $g_{bond}$ of the gradient of the link atom will be projected onto $Y_{real}$, giving rise to diagonal matrices with values of $1 - g_{bond}$ and $g_{bond}$ on the diagonal for the corresponding blocks in the Jacobian matrix, respectively.

As the positions of the link atoms depend on the positions of the *real* atoms, no additional degrees of freedom are thus introduced from the perspective of *real*.

### 3.3.2 Choosing the model region

As for what to include in the *model* region and where to cut the bonds in a macromolecular QM/MM setting, below is a non-exhaustive list with some general guidelines (Senn and Thiel, 2009; Chung et al., 2015; Ryde, 2016; Clemente et al., 2023):

- Identify the region of interest and include all atoms directly involved in the process of interest.
- Avoid cutting polar bonds. Instead, look for $sp^3$ hybridised C–C bonds to cut some distance away from the assumed process of interest.
- Never cut bonds in conjugated systems, as this will disrupt the conjugation and lead to inaccurate results.
- When treating a metal center, include the metal ion(s) and at least its complete first coordination sphere in the model region.
- Move link atoms as far away as possible from the process of interest.
- Ensure that the *model* region is charge neutral, or at least not highly charged, to avoid artefacts from long-range electrostatic interactions.
- Include residues that can have a significant influence on the electronic structure of the *model* region, such as charged residues or residues forming hydrogen bonds with the proposed *model* region.

# Chapter 4

# Quantum refinement

Quantum refinement (QR) is a powerful approach for improving macromolecular structure refinement by integrating QM calculations into the refinement process. The method was originally developed by Ryde and co-workers in 2002 and the idea is to replace the traditional geometric MM-based restraints used in macromolecular refinement with more accurate physics-based QM calculations for a region of interest in a macromolecule. This allows for detailed insights into specific regions of interest, such as active sites or ligand-binding pockets. QR has since then become an umbrella term for several different implementations and variations of the original method. In this chapter, I will give an overview of the method, as well as describe the specific implementation used in this thesis.

## 4.1 Force fields in refinement

In order for force fields to provide useful prior information to the macromolecular refinement target function (equation 2.31), they need to be able to accurately describe the geometry of the system. However, in the macromolecular refinement setting, the Coulomb term in equation 3.43 is typically omitted owing to the difficulty in assigning partial charges to atoms in a macromolecule, as well as the fact that the dielectric constant inside a protein is not well defined. Additionally, hydrogen atoms are often not included in the model, as they are typically not discernable at the resolutions commonly achieved in macromolecular structure determination, making it difficult to properly account for their electrostatic interactions in the form of hydrogen bonds.

Furthermore, the attractive part of the Lennard–Jones potential is also often omitted, im-

plying that the van der Waals interactions are purely repulsive and mainly serve the purpose to avoid steric clashes. Is is also worth noting that the commonly used force fields used in macromolecular refinement, at least originally, were derived from small-molecule crystal structures and thus do not necessarily properly account for the flexibility of macromolecules. Moreover, force fields used in macromolecular refinement are usually also formulated in a statistical manner rather than from first principles (Engh and Huber, 1991; Mackerell, 2004; Engh and Huber, 2012).

Amino acid residues and nucleotides are typically accurately described by most force fields used in macromolecular refinement (Kleywegt and Jones, 1998). However, the situation is not as good for ligands and other non-standard residues, for which the quality of the parametrisations can vary significantly between different force fields. Owing to the vastness of the chemical space, there is also another challenge in the form of parameter coverage, i.e. not all ligands or unusual chemical species are covered by a given force field or the parametrisation is less accurate than for standard residues (Merz, 2014; Kleywegt and Jones, 1998; Nilsson et al., 2003; Mackerell, 2004). This is especially problematic for metals and metal-containing compounds, for which it is very difficult to set up an accurate general force field (as the geometry depends on the element, oxidation state, spin state as well as the nature of all first-sphere ligands), making it difficult to accurately model metalloproteins using MM (Hu and Ryde, 2011). For example, the most recent update to `GeoStd`, used by `phenix.refine` and `phenix.real_space_refine`, contains approximately 37000 unique restraints for ligands, none of which cover metal ions or any metal-containing compounds (Afonine et al., 2018b; Liebschner et al., 2019; Moriarty et al., 2025).

Often it is then up to the user to provide parametrisations for their systems, through tools such as `phenix.elbow` or `AceDRG`, or by generating restraints *in situ* through tools such as `MetalCoord` or `QMR` (Moriarty et al., 2009; Long et al., 2017; Liebschner et al., 2023; Babai et al., 2024). However, neither of these tools make use of the Hessian matrix of the potential energy surface (PES), which would allow for determination of the force constants in equations 3.42 and 3.43. Instead, they provide calculated equilibrium values only.

## 4.2 Quantum refinement

The problem with the use of force fields in macromolecular refinement can be alleviated through the use of *in situ* QM calculations, which allow for a more accurate description of the electronic environment and can capture effects that are not well-represented by classical force fields. With QM calculations, restraints are generated on-the-fly during the refinement process, thus ensuring that they are always accurate and up-to-date with the current geometry of the system. The tradeoff is the increased computational cost, as QM calcula-

tions are significantly more expensive than parametrised restraint calculations.

An approach of this kind was first presented by Ryde and co-workers in 2002 for XRD, through the interface `ComQum-X`, where the method was named "quantum refinement" (Ryde et al., 2002). In `ComQum-X`, the restraint term, $T_{restraints}$ in equation 2.31, is augmented to include a QM term for a small, but interesting, part of the macromolecule, through a subtractive hybrid scheme, using a hydrogen link atom approach, i.e.

$$T_{restraints} = (w_{QM} E_{QM,model} + T_{low,real} - T_{low,model}), \tag{4.1}$$

where $E_{QM,model}$ is the energy of the model region calculated with a QM method, $T_{low,real}$ is the standard lower-level restraint term for the entire system, $T_{low,model}$ is the corresponding lower-level restraint term of the *model* region, in a manner similar to equation 3.45. In `ComQum-X`, which interfaces the crystallographic software `CNS` with the QM software `Turbomole`, $T_{low,real}$ and $T_{low,model}$ are calculated by `CNS` using the Engh & Huber parametrisation (Engh and Huber, 1991; Brünger et al., 1998; Furche et al., 2014). As QM calculations report physical energies, whereas the Engh & Huber parametrisation is statistical, an additional weight factor for the QM term, $w_{QM}$, needs to be introduced. In the original implementation, $w_{QM}$ was set to 3.

The `ComQum-X` interface was originally developed for XRD and has since been extended to also work for NMR (`ComQum-N`), extended X-ray absorption fine structure (`ComQum-EXAFS`), neutron crystallography (`ComQum-U`), joint X-ray and neutron crystallography (`ComQum-UX`) as well as two- and four-layer QM for XRD (`ComQum-X-2QM` and `ComQum-X-4QM`) (Hsiao et al., 2005, 2006; Ryde et al., 2007; Caldararu et al., 2019; Cao and Ryde, 2020; Cao et al., 2020; Cirri et al., 2022). The `ComQum` series of interfaces have been used successfully to locally improve the geometry of metal sites in macromolecules, determine protonation states of metal-bound ligands, determine oxidation states of metal sites, detect changes in oxidation states of metals during data collection as well as identifying unknown ligands, showcasing the versatility and success of the method (Ryde and Nilsson, 2003; Nilsson and Ryde, 2004; Rulísek and Ryde, 2006; Söderhjelm and Ryde, 2006; Hersleth and Andersson, 2011; Bergmann et al., 2021a).

Variants of QR have been implemented by severals other groups as well. Merz and coworkers developed a linear-scaling version of QR in 2005, where semiempirical QM (SQM) calculations are employed for the entire system for the restraints term in equation 2.31 (implying that $T_{restraints}$ is simply $E_{QM,real}$) through a divide-and-conquer approach, with `CNS` being used for the $T_{data}$ term (Yu et al., 2005). Thiel and co-workers implemented QR in 2010 in `ChemShell`, as an interface between `CNS`, `Turbomole` and `DL_POLY`, where the force field this time also includes electrostatics (Hsiao et al., 2010; Lu et al., 2019; London et al., 2025). Chung and coworkers have implemented several variants of QR in a multiscale ONIOM QM/MM framework (equation 3.46), where they combine coupled

cluster, SQM, MM and machine learning potentials for the restraints term, again using CNS for the $T_{data}$ term (Yan et al., 2021, 2024).

For the PHENIX and REFMAC refinement software, a QR implementation is available through the commercial DivCon plugin, which is a linear-scaling QR implementation at the SQM level of theory, where several steps in the setup and refinement process have been automated, making it more user-friendly (Murshudov et al., 2011; Borbulevych et al., 2014, 2018; Liebschner et al., 2019). Additionally, in 2017 an ambitious project was started, called Q|R, with the goal to develop a QR implementation based on open-source software, primarily the cctbx library (which is also the basis for PHENIX), where the entire restraints term is calculated using QM (Grosse-Kunstleve et al., 2002; Grosse-Kunstleve and Adams, 2002; Zheng et al., 2017c). A secondary goal for Q|R is to use best practises in software development to create a robust and user-friendly implementation of QR, with sustainability in mind. In Q|R, a fragment based divide-and-conquer approach is used, where the entire system is divided into overlapping fragments, which arecalculated separately with QM. The total gradient is then obtained by combining the results from the fragments (Zheng et al., 2017b; Wang et al., 2023). Q|R has also been extended to handle symmetry interactions, as well as work for cryo-EM (Zheng et al., 2020; Wang et al., 2020). A recent development of the Q|R project is the addition of a machine learning potential for the restraints, trained on DFT QM calculations, in the form of the AQuaRef package (Zubatyuk et al., 2025).

## 4.3 QRef

A significant portion of the work during this thesis was spent on developing a new QR implementation, called QRef, which is an interface between the macromolecular refinement software PHENIX and the QM software ORCA (Neese, 2012; Liebschner et al., 2019; Neese, 2025). QRef was developed from scratch in Python, using a subtractive hybrid scheme, with hydrogen link atoms.

With the wealth of QR implementations already available, the question is then why develop yet another implementation? The answer is multifaceted:

- **Metals:** Our research group is particularly interested in metalloproteins, where some of the existing implementations of QR are not ideal. Either they do not support heavier elements at all or are treated at a, in our opinion, too low level of theory.
- **Hybrid methods:** In our experience it is sufficient to treat only a small part of the system with QM and spend the computational efforts where they matter the most. For a few of the existing implementations, there is no option to chose which part of the system to treat with QM.

- **Control:** Being able to chose the level of theory for the QM calculations is important, as different systems have different requirements. Many of the existing implementations are either limited to SQM methods or a specific set of DFT functionals and basis sets. In `QRef`, the user can chose any method and basis set available in `ORCA`, allowing for greater control and flexibility.

- **Open-source:** Some of the existing implementations are either commercial or not open-source, which limits their accessibility and usability for the wider scientific community. `QRef` is fully open-source, released under a BSD-3-Clause license and freely available on GitHub at `https://github.com/krlun/QRef`.

- **Flexibility:** Many of the existing implementations are tightly coupled to specific software packages or QM methods, which can limit their flexibility and applicability to different systems. While `ORCA` was chosen as the QM software for `QRef`, the `QRef` code is modular and can easily be extended to support other QM software in the future.

- **User-friendliness:** Many of the existing QR implementations are not very user-friendly, requiring significant manual setup and expertise to use effectively. `QRef` was developed with at least an intention to be decently user-friendly, but there is still room for improvement and automation of the setup process.

- **Sustainability:** Some of the existing implementations of QR are not particularly maintained or updated, which can lead to compatibility issues with newer software versions or lack of support for new features. `QRef` is actively maintained and updated and has been tested to work with `PHENIX` versions 1.20.1-4487, 1.21-5207, 1.21.1-5286 and 1.21.2-5419, as well as `ORCA` versions 5.0.3, 5.0.4, 6.0.1 and 6.1.0.

While the `ComQum` series of interfaces, also developed by our research group, fulfill many of these criteria, they do, like many other QR implementations, rely on `CNS` for the $T_{data}$ term. As `CNS` has not been updated since 2010 (Brünger et al., 2010), thus lacking many modern features found in more recent refinement software packages, this was an additional driving force for implementing `QRef` in `PHENIX`, moving over to more modern refinement software. Additionally, `CNS` only has support for reciprocal space refinement for XRD and ND, whereas `PHENIX` also supports real-space refinement, as well as ED. While a quite trivial shift, changing the QM software from `Turbomole` to `ORCA` was motivated by the fact that `Turbomole` is commercial software, whereas `ORCA` is free for academic use.

### 4.3.1 Implementation details

The geometrical restraint target in `QRef` is implemented as

$$T_{restraints}(\mathbf{R}_{real}) = \sum_i \left( w_{QM} E_{QM,i}(\mathbf{R}_{model_i}) - T_{low,i}(\mathbf{R}_{model_i}) \right) + T_{low,real}(\mathbf{R}_{real}), \quad (4.2)$$

where the sum is over all *model* regions and thus allows an arbitrary number of QM regions, which can be disjunct or overlapping. The QM terms are calculated using `ORCA`, while the lower-level terms are calculated using the `PHENIX` internal restraint engine in the form of `cctbx`. Like in equation 4.1, a weight factor for the QM term, $w_{QM}$, is introduced to balance the contribution from the QM calculations with the lower-level terms. From comparing the restraints in `PHENIX` with the `AMBER` force field `ff14SB` (Maier et al., 2015), a value of 7.5 mol/kcal for $w_{QM}$ was found to be appropriate. Subsequent internal testing (unpublished) through comparison of gradient norms between $E_{QM}$ and $T_{low}$ for perturbed amino acid residues in a manner similar to equation 2.47 has indicated that this value is reasonable, but should perhaps be slightly lower, approximately 6–7 mol/kcal.

Owing to limitations in the restraint handling of `cctbx`, the (*low, model*) representation is not capped with hydrogen link atoms, instead the link atoms are kept at their original position with their original element type, i.e. (*low, model*) is a truncation of (*low, real*). This is completely valid and leads to an exact cancellation of the low-level terms between $T_{low,real}$ and $T_{low,model}$ in equation 4.2, However, with hydrogen link atoms in (*low, model*), there is a possibility to correct for errors introduced by the link atoms in $T_{high,model}$, although the gain has been shown to be minimal in real applications (Cao and Ryde, 2018).

The corresponding gradient for $T_{restraints}$ in equation 4.2 is then given by

$$\begin{aligned} \nabla T_{restraints}(\mathbf{R}_{real}) = \sum_i & \left( w_{QM} \nabla E_{QM,i}(\mathbf{R}_{model_i}) J(\mathbf{R}_{model_i}; \mathbf{R}_{real}) - \nabla T_{low,i}(\mathbf{R}_{model_i}) \right) \\ & + \nabla T_{low,real}(\mathbf{R}_{real}), \end{aligned} \quad (4.3)$$

where the gradients of the QM terms are projected onto the *real* region through the Jacobian matrix as described following equation 3.49.

The `QRef` interface is inserted into the `phenix.refine` and `phenix.real_space_refine` programs through modification of the constructor of the `energies` class in `cctbx` for the geometry restraints manager, so that whenever a call to calculate a coordinate restraint target and gradient is made, these are modified according to equations 4.2 and 4.3. Additionally, so that any settings set by the user in the `phenix.refine` or `phenix.real_space_refine` input files are preserved, the `manager` class in `mmtbx` (the macromolecular

toolbox part of `cctbx`) is also modified to write these settings to disk, so that they can be read by `QRef` when calculating the $T_{low,i}$ terms with the same settings as the $T_{low,real}$ term. While this ensures the same settings are used to calculate the both terms, the Conformation Dependent Library (CDL), used by `PHENIX` to provide context-dependent ideal values for bond lengths and angles for the protein backbone, does not necessarily find the same environment for the two representations, causing the gradients to not cancel properly (Moriarty et al., 2016). For this reason, we recommend that CDL be explicitly turned off when using `QRef`.

A supporting script, `qref_prep.py`, is also provided, which prepares a settings file, `qref.dat`, needed by `QRef`. Several options exist for `qref_prep.py`:

- **`-c`** or **`--cif`**: Used to specify restraint files in `CIF` format for any non-standard residues or ligands in the *model* region(s).
- **`-s`** or **`--syst1`**: Specifies the file(s) containing a definition(s) of the *model* region(s). If this option is not used, the default is to look for a file named `syst1`[1] in the current working directory.
- **`-j`** or **`--junctfactor`**: Option to provide a custom database file containing the $d_{X-L}$ values for the link atoms in equation 3.48. If this option is not used, the default is to look for a file named `junctfactor` in the current working directory.
- **`-l`** or **`--ltype`**: Used to choose the level of theory for the $d_{X-L}$ distances. If this option is not used, the default is to use TPSS/def2-SV(P).
- **`-w`** or **`--w_qm`**: Can be used to change the weight factor for the QM term(s), $w_{QM}$ in equation 4.2. If this option is not used, the default is 7.5 mol/kcal.
- **`-r`** or **`--restart`**: If this option is used, for each iteration in the refinement, the current model will be written to disk in `PDB` format, allowing the user to both inspect the progress of the refinement as well as restart the refinement from the latest iteration - useful if the refinement crashes for some reason, or if the time limit on a compute cluster is reached.
- **`-rd`** or **`--restraint_distance`**: Used to specify distance restraints between pairs of atoms in the system.
- **`-ra`** or **`--restraint_angle`**: Used to specify angle restraints between triplets of atoms in the system.

[1]The naming of the files defining the *model* regions is legacy from the `ComQum` interfaces.

- `-t` or `--transform`: Allows for symmetry transformations of selected atoms in the *model* region(s), where the transformation is given by a rotation matrix, $\mathbf{R}$, and a translation vector, $\mathbf{t}$, in Cartesian coordinates, i.e. $\mathbf{r}' = \mathbf{R}\mathbf{r} + \mathbf{t}$. This is useful if the *model* region(s) involves atoms from symmetry mates.

Additionally, `qref_prep.py` will provide a selection string for each of the model regions, as well as check that each model region consists of a single alternative location. If multiple alternative locations are found in a single model region, a warning will be issued.

If `qref.dat` is not present in the current working directory when running `phenix.refine` or `phenix.real_space_refine`, `QRef` will not be activated and a non-QR refinement will run instead.

### 4.3.2 Manual restraints

As any manually added restraints to a refinement from the perspective of the refinement engine would cancel out in the subtractive scheme, support for these must be added in the `QRef` layer for them to have any effect. The possibility to use bond length restraints has been added to `QRef` in the form of harmonic potentials, $U_{ij}$ between atoms $i$ and $j$, as

$$U_{ij} = k_{ij}(d_{ij} - d_{ij}^0)^2, \tag{4.4}$$

where $k_{ij}$ is the force constant, $d_{ij}$ is the current distance between atoms $i$ and $j$, and $d_{ij}^0$ is the target distance. The partial derivatives of the potential in equation 4.4 with respect to the positions of atoms $i$ and $j$ are given by

$$\begin{aligned} \frac{\partial U_{ij}}{\partial \mathbf{r}_i} &= \frac{\partial U_{ij}}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial \mathbf{r}_i} = 2k_{ij}(d_{ij} - d_{ij}^0)\frac{(\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_j - \mathbf{r}_i|}, \\ \frac{\partial U_{ij}}{\partial \mathbf{r}_j} &= \frac{\partial U_{ij}}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial \mathbf{r}_j} = 2k_{ij}(d_{ij} - d_{ij}^0)\frac{(\mathbf{r}_j - \mathbf{r}_i)}{|\mathbf{r}_j - \mathbf{r}_i|}, \end{aligned} \tag{4.5}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the positions of atoms $i$ and $j$, respectively.

Likewise, the possibility to use angle restraints has also been added to `QRef` through harmonic potentials, $U_{ijk}$ between atoms $i$, $j$ and $k$, where $j$ is the central atom, as

$$U_{ijk} = k_{ijk}(\theta_{ijk} - \theta_{ijk}^0)^2, \tag{4.6}$$

where $k_{ijk}$ is the force constant, $\theta_{ijk}$ is the current angle formed by atoms $i$, $j$ and $k$, and $\theta_{ijk}^0$ is the target angle. The partial derivatives of the potential in equation 4.6 with respect to

the positions of atoms $i$, $j$ and $k$ are given by

$$\begin{aligned}
\frac{\partial U_{ijk}}{\partial \mathbf{r}_i} &= \frac{\partial U_{ijk}}{\partial \theta_{ijk}} \frac{\partial \theta_{ijk}}{\partial \mathbf{r}_i} = \frac{2k_{ijk}(\theta_{ijk} - \theta^0_{ijk})}{|\mathbf{r}_j - \mathbf{r}_i| \sin\theta_{ijk}} \left( \frac{\mathbf{r}_j - \mathbf{r}_k}{|\mathbf{r}_j - \mathbf{r}_k|} - \cos\theta_{ijk} \frac{\mathbf{r}_j - \mathbf{r}_i}{|\mathbf{r}_j - \mathbf{r}_i|} \right), \\
\frac{\partial U_{ijk}}{\partial \mathbf{r}_k} &= \frac{\partial U_{ijk}}{\partial \theta_{ijk}} \frac{\partial \theta_{ijk}}{\partial \mathbf{r}_k} = \frac{2k_{ijk}(\theta_{ijk} - \theta^0_{ijk})}{|\mathbf{r}_j - \mathbf{r}_k| \sin\theta_{ijk}} \left( \frac{\mathbf{r}_j - \mathbf{r}_i}{|\mathbf{r}_j - \mathbf{r}_i|} - \cos\theta_{ijk} \frac{\mathbf{r}_j - \mathbf{r}_k}{|\mathbf{r}_j - \mathbf{r}_k|} \right), \\
\frac{\partial U_{ijk}}{\partial \mathbf{r}_j} &= - \left( \frac{\partial U_{ijk}}{\partial \mathbf{r}_i} + \frac{\partial U_{ijk}}{\partial \mathbf{r}_k} \right),
\end{aligned} \tag{4.7}$$

where $\mathbf{r}_i$, $\mathbf{r}_j$ and $\mathbf{r}_k$ are the positions of atoms $i$, $j$ and $k$, respectively. When using bond length and angle restraints, the values of the corresponding potentials, $U_{ij}$ and $U_{ijk}$, are added to $T_{restraints}$, whereas the corresponding partial derivatives are added to $\nabla T_{restraints}$.

From experience, we have found that reasonable values for $k_{ij}$ and $k_{ijk}$ are approximately 2500 $\text{Å}^{-2}$ and 2500 $\text{rad}^{-2}$, respectively, in order to achieve adherence to the bond length and angle restraints, without overly constraining the system.

While more restraint types (such as proper and improper dihedral angles) could be added in a similar manner, these have not been implemented in `QRef` yet, as we have not found a need for them so far.

### 4.3.3 Workflow

The general workflow for a QR calculation using `QRef` is as follows:

1. Obtain a starting structure and the corresponding experimental data to be used in the refinement.
2. Identify the region(s) of interest to be treated with QM, adhering to the guidelines in chapter 3, section 3.3.2.
3. Decide on the level of theory and basis set to be used for the QM calculations, as well as charge and multiplicity. We have used the DFT TPSS functional with the def2-SV(P) basis set. While a larger basis set could be motivated from QM accuracy perspective, unless the data is of very high quality (better than 1.1-1.2 Å), the gain in accuracy is likely negligible compared to the increased computational cost. Additionally, we recommend using DFT–D4, to account for dispersion interactions (Caldeweyher et al., 2019). The settings for the QM calculations should be specified in files `qm_i.inp`, where `i` is the index of the model region, starting from 1.

4. Prepare the input structure, ensuring that it is properly protonated. A QM calculation will calculate what it is given (which also allows for hypothesis testing). We recommend using `phenix.ready_set`.

5. Generate restraint files for any non-standard residues or ligands in the system, if needed. We recommend using `phenix.elbow` for this step. In a subtractive scheme, restraints in the *model* region will cancel out; thus these restraints do not need to be of high quality. However, they do need to exist. If restraint files are needed, make sure that the same restraint files are used by both `PHENIX` and `QRef`.

6. Define the *model* region(s), i.e. the part of the structure to be treated with QM. This is identified by the `SERIAL` field in the `PDB` file. In order to maintain a consistent atom ordering with that used by `PHENIX` internally, a script `sort_pdb.py` is also provided with `QRef`.

7. Set up the `PHENIX` refinement input file, specifying any settings for the refinement as well as the experimental data to be used. We recommend the following options to be set for reciprocal space refinement:

    - `refinement.pdb_interpretation.restraints_library.cdl = False` - turn off CDL
    - `refinement.pdb_interpretation.restraints_library.mcl = False` - turn off the metal coordination library
    - `refinement.pdb_interpretation.restraints_library.cis_pro_eh99 = False` - turn off the cis peptide bond library
    - `refinement.pdb_interpretation.secondary_structure.enabled = False` - turn off secondary structure restraints, as these will not be valid in the *model* region(s)
    - `refinement.pdb_interpretation.sort_atoms = False` - turn off automatic sorting of atoms, as this can change the atom order in the *model* region(s) compared to the input model, causing problems for `QRef`
    - `refinement.pdb_interpretation.flip_symmetric_amino_acids = False` - turn off automatic flipping of symmetric amino acids
    - `refinement.refine.strategy = *individual_sites individual_sites_real_space rigid_body *individual_adp group_adp tls occupancies group_anomalous` - use individual sites refinement and ADP refinement; QR is typically run at the end of structure determination, thus the other strategies do not make sense to use
    - `refinement.refine.sites.individual = <reciprocal selection string>` - specify which atoms to refine (use the selection strings

provided by `qref_prep.py`); we recommend only refining atoms in the *model* region(s) and any residue connected to it/them

- `refinement.hydrogens.refine = *individual riding Automatic` - with QR, hydrogens can be refined individually, as the QM calculations will provide a more accurate description of their positions than riding hydrogen atoms
- `refinement.main.nqh_flips = False` - turn off automatic flipping of Asn, Gln and His side chains, as this can cause convergence issues in the QM SCF procedure

In the case of real space refinement, we recommend the following options:

- `refinement.run = *minimization_global rigid_body local_grid_search morphing simulated_annealing adp occupancy nqh_flips` - only do coordinate refinement
- `pdb_interpretation.restraints_library.cdl = False` - turn off CDL
- `pdb_interpretation.restraints_library.mcl = False` - turn off the metal coordination library
- `pdb_interpretation.restraints_library.cis_pro_eh99 = False` - turn off the cis peptide bond library
- `pdb_interpretation.flip_symmetric_amino_acids = False` - turn off automatic flipping of symmetric amino acids
- `pdb_interpretation.sort_atoms = False` - turn off automatic sorting of atoms
- `pdb_interpretation.secondary_structure = False` - turn off secondary structure restraints
- `pdb_interpretation.reference_coordinate_restraints.enabled = True` - turn on reference coordinate restraints, as this is the only way to restrain atoms in real space refinement
- `pdb_interpretation.reference_coordinate_restraints.selection = <real space selection string>` - specify which atoms to restrain (use the selection strings provided by `qref_prep.py`)
- `pdb_interpretation.reference_coordinate_restraints.sigma = 0.01` - set the strength of the reference coordinate restraints; a value of 0.01 is typically reasonable
- `pdb_interpretation.ramachandran_plot_restraints.enabled = False` - turn off Ramachandran plot restraints, as these will not be valid in the *model* region(s)

8. Run `qref_prep.py` to generate the `qref.dat` settings file.
9. Run the refinement using `phenix.refine` or `phenix.real_space_refine`, which will automatically call `QRef` to calculate the geometric restraint term and gradient during each refinement iteration.
10. Analyse the results, including the refined structure, validation metrics, and any changes in the geometry.

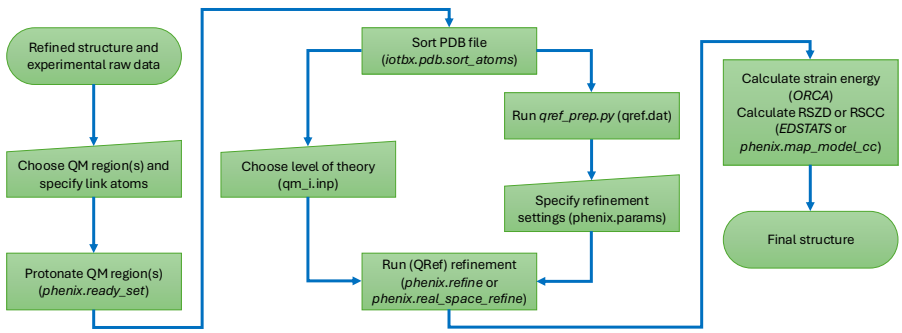



**Figure 4.1:** Schematic overview of the `QRef` workflow.

A more detailed description on how to install `QRef`, setup and run QR calculations using `QRef` can be found in the `QRef` documentation at `https://github.com/krlun/QRef`. Additionally, several examples of QR setups using `QRef` can be found in the `examples` directory of the `QRef` GitHub repository.

## 4.4 Validation metrics

As the goal of quantum refinement is to either improve a given model, or to perform hypothesis testing, through the implicit effects of the QM treatment, as to elucidate what is the most likely model that explains the experimental data, it becomes important to have metrics to validate the results, which in turn helps ensure that the result is suitable for further uses (Read et al., 2011).

In the reciprocal setting, the *R*-factor is commonly used to assess the global quality of a model, given by

$$R = \frac{\sum_{hkl} ||F_{hkl}^{obs}| - |F_{hkl}^{model}||}{\sum_{hkl} |F_{hkl}^{obs}|}. \tag{4.8}$$

If the model perfectly explains the experimental data, the *R*-factor will be zero. With the large number of parameters that can be refined for a model, there is a significant risk of overfitting the model to the data, which in turn can lead to a low *R*-factor even for a poor model. In order to alleviate this problem, the free *R*-factor, $R_{free}$, is often used as a more robust metric, where a subset of the reflections (typically around 5–10%) are set aside and not used during the refinement process. This subset is then used to calculate $R_{free}$, which thus provides an unbiased estimate of how well the model explains the experimental data (Brünger, 1992). Commonly, $R_{free}$ is expected to be higher than the *R*-factor and a large difference (bigger than $\approx 0.03$) between the two can indicate overfitting (Proteopedia, 2025).

In quantum refinement, at least in the setting used in this thesis (where the focus is on a small part of the model), local quality measures are of interest rather than global *R*-factors. Two such metrics are the real-space *R* factor (RSR) and the real-space Z-difference (RSZD) score, which can be calculated for specific regions of the model, hence allowing for local assessment of the model quality. The RSR factor is given by

$$RSR = \frac{\sum_i |\rho_{obs,i} - \rho_{calc,i}|}{\sum_i |\rho_{obs,i} + \rho_{calc,i}|}, \tag{4.9}$$

where $\rho_{obs,i}$ and $\rho_{calc,i}$ are the observed and calculated densities, respectively, at voxel $i$. The RSR can be truncated to be calculated only for a specific region of interest, such as around a ligand or residue. While several versions of the RSR exist, they all have in common that they are correlated to both the accuracy and precision of the model, which is not ideal (Jones et al., 1991; Winn et al., 2011; Tickle, 2012).

The RSZD score instead uses a $\chi$-squared approach to calculate the difference density, $\Delta\rho = \rho_{obs} - \rho_{calc}$, in order to assess the local quality of a model in real space and is given by

$$RSZD = \frac{\Delta\rho}{\sigma(\Delta\rho)} \tag{4.10}$$

where $\sigma(\Delta\rho)$ is the standard deviation of the difference density in that region. The RSZD score quantifies the significance of positive or negative difference densities in a given region (e.g. around a ligand or residue) by normalising the difference density to the expected local density noise ($\sigma(\Delta\rho)$), giving a Z-score–style metric. RSZD is therefore an accuracy metric, rather than a precision metric, of how well the model explains the experimental data, instead of how well the model fits the data. Large |RSZD| values indicate significant unexplained density (model inaccuracies or missing features), while values near zero show

that the model adequately explains the data. An RSZD score with an absolute value above three is typically considered a significant outlier and suggests rejection of that part of the model (Tickle, 2012).

In the real-space setting, there are no experimental structure factor amplitudes to compare against as in equation 4.8, thus other metrics than reciprocal *R*-factors must be used. One commonly used metric is the real-space correlation coefficient (RSCC), which measures the correlation between the density calculated from the model and the experimental density. The RSCC is given by

$$RSCC = \frac{\sum_i (\rho_{obs,i} - \langle \rho_{obs} \rangle)(\rho_{calc,i} - \langle \rho_{calc} \rangle)}{\sqrt{\sum_i (\rho_{obs,i} - \langle \rho_{obs} \rangle)^2 \sum_i (\rho_{calc,i} - \langle \rho_{calc} \rangle)^2}}, \tag{4.11}$$

where $\rho_{obs,i}$ and $\rho_{calc,i}$ are the observed and calculated densities at voxel $i$, respectively, and $\langle \rho_{obs} \rangle$ and $\langle \rho_{calc} \rangle$ are the mean observed and calculated densities, respectively, summed over all voxels $i$ in a given region. The RSCC ranges from –1 to 1, where 1 indicates perfect correlation, 0 indicates no correlation, and –1 indicates perfect anti-correlation. A high RSCC value (typically above 0.8) indicates that the model fits well to the experimental density, while a low RSCC value indicates a poor fit (Urzhumtsev et al., 2014; Joseph et al., 2017; Afonine et al., 2018a). The RSCC can be calculated for specific regions of the density, allowing for local assessment of the model quality.

Equations 2.13 and 2.14 show that the structure factors and the density are Fourier transforms of each other and consequently RSR, RSZD and RSCC are in theory possible to calculate for models obtained from both reciprocal and real-space data. However, estimating $\sigma(\Delta\rho)$ in equation 4.10 for cryo-EM densities is a non-trivial task. Because of this reason, RSZD is typically only used in the reciprocal-space setting. In real-space and cryo-EM data sets, where resolution currently tends to be lower, RSCC offers a good alternative.

Additionally, visual inspection of the resulting model and the fit to the density is always recommended. For this purpose, the likelihood-weighted $2mF_o - DF_c$ and $mF_o - DF_c$ maps can be used in reciprocal space, where $m$ and $D$ are the figure of merit and the overall scale factor, respectively, and $F_o$ and $F_c$ are the observed and calculated structure factors, respectively (Luzzati, 1952; Urzhumtsev et al., 1996; Lunin et al., 2002). The $2mF_o - DF_c$ map is a weighted version of the traditional $2F_o - F_c$ map, which is less noisy and thus easier to interpret, whereas the $mFo - DF_c$ map is a weighted version of the traditional $F_o - F_c$ map, which highlights differences between the model and the experimental data (Rhodes, 2006; Rupp, 2009). In real space, where the density is not model biased, the experimental density map can be used directly for visual inspection of the model fit to the data. The resulting model itself (in particular its geometry) should also be examined in terms of its chemical plausibility.

Furthermore, the QM energy (which is obtained in a sense for free during QR) of the final refined model can be used as an additional local metric for the QM system itself in the form of strain energies relative to some sort of reference structure.

A few different options can be considered for the definition of the strain energy, $E_{strain}$. The intuitively most obvious option (used in the first QR studies) is to simply do a free geometry optimisation of the QM region and compare the obtained energy from this optimised structure with the energy of the QM region in the QR structure (Ryde et al., 2002). This works well for an isolated molecule or a small rigid QM region (e.g. a metal site with only the first-sphere ligands). However, for larger QM regions with some weakly connected moieties, such a reference structure becomes increasingly problematic, as groups can move around widely during the geometry optimisations, forming interactions (or even new chemical species) that are not possible for the native structure. Therefore, some sort of restraints or constraints normally need to be used during the geometry optimisation to keep the reference structure in the same structural domain as in the protein.

One way is to run two different refinements, the first with a $w_{data}$ factor in equation 2.31 greater than zero (i.e. the normal QR calculation), followed by another refinement of the resulting model from the first refinement, this time with $w_{data} = 0$ (i.e. a geometry optimisation based on the refinement restraints only, essentially a QM/MM optimisation). The strain energy is then defined as

$$\Delta E_{strain} = E_{QM}(w_{data} > 0) - E_{QM}(w_{data} = 0), \tag{4.12}$$

where $E_{QM}(w_{data} > 0)$ and $E_{QM}(w_{data} = 0)$ are the QM energies of the final models obtained from the two refinements, respectively. This is the approach used in later studies with `ComQum-X`, but initial tests with `QRef` gave rather noisy strain energies. This is possibly an artifact due to the fact that in `ComQum-X` refinement is performed from the *model* perspective, whereas in `QRef` refinement is done from the *real* perspective.

A second option for the strain energy, which has generally been used in this thesis, is to take the final model obtained from a refinement with $w_{data} > 0$ and then perform a QM optimisation of the QM system alone, with frozen link atoms. The strain energy in this case is then defined as

$$\Delta E_{strain} = E_{QM}(w_{data} > 0) - E_{QM,opt}, \tag{4.13}$$

where $E_{QM,opt}$ is the QM energy of the optimised structure. This definition of the strain energy reduces the problem of ending up in different local minima, but the reference state will depend on the positions of the frozen link atoms, which may differ slightly between different hypotheses.

In the case of performing scans of $w_{data}$ (e.g. in order to find a suitable value of $w_{data}$ to use in hypotheses testing), it is often best to use the same reference structure for all QR structures, in order to make the strain energies more stable.

A low strain energy indicates that the prior agrees with the likelihood, i.e. the proposed model is in agreement with both the experimental data and the QM calculations, and increasing strain energies indicate a growing discrepancy. It should also be noted that QM energies depend on the size and net charge of the QM system. Therefore, strain energies should only be compared between models with the same atoms and net charge. For hypotheses that differ in terms of atoms and/or net charge, a thorough assessment is warranted when comparing strain energies (Bergmann et al., 2022).

Typically, biasing the the refinement target towards the experimental data improves the RSR, RSZD and RSCC scores, while simultaneously increasing the strain energy, and vice versa. In turn, this also allows for selection of the $w_{data}$ factor in equation 2.31 in regards to hypotheses testing, which is further discussed in **Paper II**.

# Chapter 5

# Studied proteins

## 5.1 Nitrogenase

All life as we know it depends on nitrogen, as it is a key component of amino acids and nucleotides (Aczel, 2019). While nitrogen gas ($N_2$) makes up approximately 78% of the Earth's atmosphere, most organisms are unable to use it directly because of the strong triple bond in $N_2$ and must hence rely on fixed forms of nitrogen, such as ammonia ($NH_3$) or nitrate ($NO_3^-$) (Erisman et al., 2008; Hoffman et al., 2014). Converting $N_2$ into $NH_3$ is known as "nitrogen fixation" and can occur through biological, abiotic, as well as industrial processes (Erisman et al., 2008).

Industrially, ammonia is produced through the Haber–Bosch process, where $N_2$ and hydrogen gas ($H_2$) are combined at high temperature and pressure in the presence of a metal catalyst in order to produce $NH_3$ (Smil, 2004). Currently, the Haber–Bosch process consumes about 1% of the world's energy production, as well as being responsible for approximately 1.4% of the world's $CO_2$ emissions (Capdevila-Cortada, 2019; Zhang et al., 2020). There is thus a strong interest in understanding biological nitrogen fixation, which occurs at ambient conditions, as it could lead to more sustainable methods for producing fixed nitrogen.

The only known enzymes that can catalyse the reduction of $N_2$ to $NH_3$ are nitrogenases, which are found in certain bacteria and archaea, known as *diazotrophs* (Zhao et al., 2006). Three variants of nitrogenase are known, which differ in the metal content of their active site cofactors: molybdenum (Mo)-nitrogenase, vanadium (V)-nitrogenase and iron (Fe)-only nitrogenase. Mo-nitrogenase is the most well-studied and also the most efficient of the three, with V-nitrogenase and Fe-only nitrogenase being less efficient and only expressed under certain conditions, such as when Mo is scarce (Seefeldt et al., 2020).

The catalysed reaction for the respective nitrogenases can be approximately summarised as (Harris et al., 2019):

$$\begin{aligned}
\text{Mo}: &\quad \mathrm{N_2} + 8\,\mathrm{H^+} + 8\,\mathrm{e^-} + 16\,\mathrm{ATP} \longrightarrow 2\,\mathrm{NH_3} + \mathrm{H_2} + 16\,\mathrm{ADP} + 16\,\mathrm{P_i} \\
\text{V}: &\quad \mathrm{N_2} + 18\,\mathrm{H^+} + 18\,\mathrm{e^-} + 36\,\mathrm{ATP} \longrightarrow 2\,\mathrm{NH_3} + 6\,\mathrm{H_2} + 36\,\mathrm{ADP} + 36\,\mathrm{P_i} \\
\text{Fe}: &\quad \mathrm{N_2} + 20\,\mathrm{H^+} + 20\,\mathrm{e^-} + 40\,\mathrm{ATP} \longrightarrow 2\,\mathrm{NH_3} + 8\,\mathrm{H_2} + 40\,\mathrm{ADP} + 40\,\mathrm{P_i}.
\end{aligned} \tag{5.1}$$

Interestingly, all three nitrogenases also produce $H_2$ as a by-product, which could be harnessed as a potential source of renewable energy (Zhang et al., 2020). All three proteins are heterotetramers and the active site has been identified to be a complex metal cluster consisting of Mo/V/Fe/$Fe_7S_9C$, which in all versions is coordinated by at least a cysteine residue, a histidine residue and a homocitrate ligand.

While the reaction mechanism has been studied extensively, both experimentally and computationally, many details remain unclear. It has been suggested that Mo-nitrogenase operates through a series of eight intermediate states, $E_1$ to $E_8$, where the subscript indicates the number of electrons and protons that have been added to the active site for each step, together with $E_0$ as the resting state, known as the "Lowe–Thorneley" kinetic model (Burgess and Lowe, 1996; Harris et al., 2019).

In this thesis, Fe-nitrogenase has been studied in **Papers I** and **II**, whereas V-nitrogenase has been studied in **Paper II**.

## 5.2 Manganese superoxide dismutase

Superoxide dismutases (SODs) are enzymes that catalyses the dismutation of superoxide ($O_2^{\bullet -}$) into oxygen ($O_2$) and hydrogen peroxide ($H_2O_2$). The highly reactive and poisonous superoxide molecule is a by-product of oxygen metabolism in living organisms. If not regulated properly, superoxide can cause significant damage to cells through oxidative stress. Thus, SODs play a crucial role in protecting living cells (Zheng et al., 2023).

Superoxide dismutases are present in almost all living organisms and the genes encoding for SODs are highly conserved between species, indicating their essential role in cellular defense mechanisms (Tian et al., 2021). In humans, three types of SODs are present, namely cytosolic Cu/Zn-SOD (SOD1), mitochondrial MnSOD (SOD2) and extracellular Cu/Zn-SOD (SOD3) (Zheng et al., 2023). In this thesis, MnSOD has been studied in **Papers II** and **V**.

MnSOD has the overall reaction

$$\begin{aligned} \text{Ox:} \quad & \mathrm{Mn_2^+ + O_2^{\bullet -} \rightarrow Mn_3^+ + O_2} \\ \text{Red:} \quad & \mathrm{Mn_3^+ + O_2^{\bullet -} + 2\,H^+ \rightarrow Mn_2^+ + H_2O_2}, \end{aligned} \tag{5.2}$$

where the Mn ion cycles between the +2 and +3 oxidation states during the catalytic cycle (Azadmanesh and Borgstahl, 2018; Zheng et al., 2023; Azadmanesh et al., 2024). The active site of MnSOD consists of a Mn ion coordinated by three histidine residues, one aspartate residue and one solvent molecule, typically water or hydroxide (Borgstahl et al., 1992, 1996). The reaction mechanism of MnSOD is unclear. Both inner- and outer-sphere mechanisms, or a combination of the two, have been suggested (Azadmanesh and Borgstahl, 2018; Srnec et al., 2009; Abreu and Cabelli, 2010; Sheng et al., 2014).

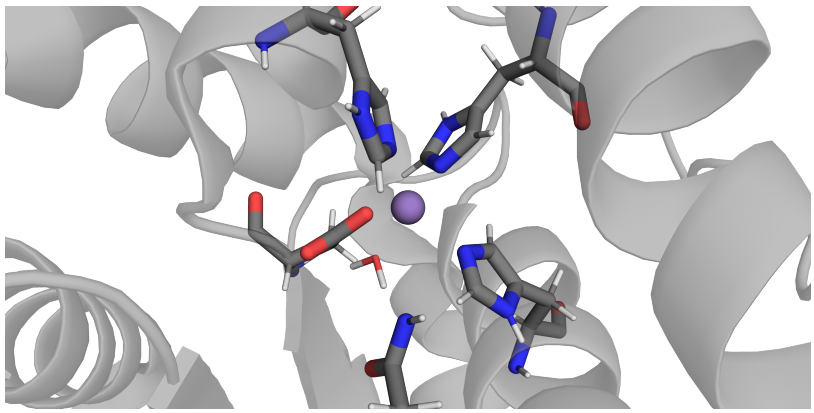


**Figure 5.1:** Active site of reduced human MnSOD expressed in *Escherichia coli* (PDB ID 7KKW) at 2.30 Å resolution, obtained via ND (Azadmanesh et al., 2021). The active site of subunit B is shown, with the Mn ion (purple) coordinated by three histidine residues, one aspartate residue, a suspicious deprotonated glutamine and a water molecule. Figure made with PyMOL (Schrödinger, 2024).

Dysfunction of MnSOD has in humans been linked to several diseases, including cancer, neurodegenerative disorders and cardiovascular diseases (Albers and Flint Beal, 2000; Valenti et al., 2004; Dhar and St Clair, 2012; Hart et al., 2015; Azadmanesh et al., 2024). With the medical implications of MnSOD, there is a strong interest in understanding its structure and function, which could lead to the development of new therapeutic strategies for diseases associated with oxidative stress.

## 5.3 Particulate methane monooxygenase

Methane monooxygenases are the only known enzymes that can catalyse the conversion of methane ($CH_4$) to methanol ($CH_3OH$) under ambient conditions (Culpepper and Rosen-

zweig, 2012). These enzymes are expressed by bacteria known as *methanotrophs*, which use methane as a source of carbon and energy (Hanson and Hanson, 1996).

This reaction is of great interest, as methane is a potent greenhouse gas and a major component of natural gas but it is hard to transport and store, while methanol is a valuable chemical feedstock and fuel (Jiang et al., 2010). Elucidating the structure and function of methane monooxygenases has consequently drawn a lot of research interest, as it could lead to the development of new catalysts for methane conversion (Ross and Rosenzweig, 2017).

Two types of methane monooxygenases are known, soluble methane monooxygenase (sMMO) and particulate methane monooxygenase (pMMO). sMMOs have a di-iron active site and are found in the cytoplasm of methanotrophs. They are rather well understood (Rosenzweig et al., 1993; Elango et al., 1997; Merkx et al., 2001; Ross and Rosenzweig, 2017). In contrast, pMMOs contain copper and are membrane-bound, which makes them inherently more difficult to study due to difficulties in the expression and crystallisation process.

The first crystal structure of pMMO was solved in 2005 at 2.8 Å resolution by Rosenzweig and coworkers and it was found to be a trimer, with each protomer consisting of three subunits, PmoA, PmoB and PmoC and each protomer was found to have three metal centers (Lieberman and Rosenzweig, 2005b). The first metal center is a copper site ($Cu_A$) in the soluble PmoB subunit, coordinated by two histidine residues. A second copper site ($Cu_B$) was also identified in the PmoB subunit, coordinated by three histidine residues. It was originally suggested from EXAFS studies that $Cu_B$ is a di-metal site and it was modelled as such in the first crystal structures (Lieberman et al., 2003; Martinho et al., 2007; Rosenzweig, 2008). However, a QR study by us showed that $Cu_B$ is much better modelled as a mono-copper site (Cao et al., 2018). This has also been supported by later crystallographic and spectroscopic studies (Ross et al., 2019; Ro et al., 2019). The third metal center, $Cu_C$, was found in the membrane-bound PmoC subunit and was originally interpreted to contain a zinc ion, ligated by two histidine residues and one aspartate residue. The zinc ion was in a later study suggested to be a crystallisation artefact, as zinc was present in the crystallisation buffer of the original study and $Cu_C$ was instead proposed to be a copper ion (Smith et al., 2011; Culpepper and Rosenzweig, 2012).

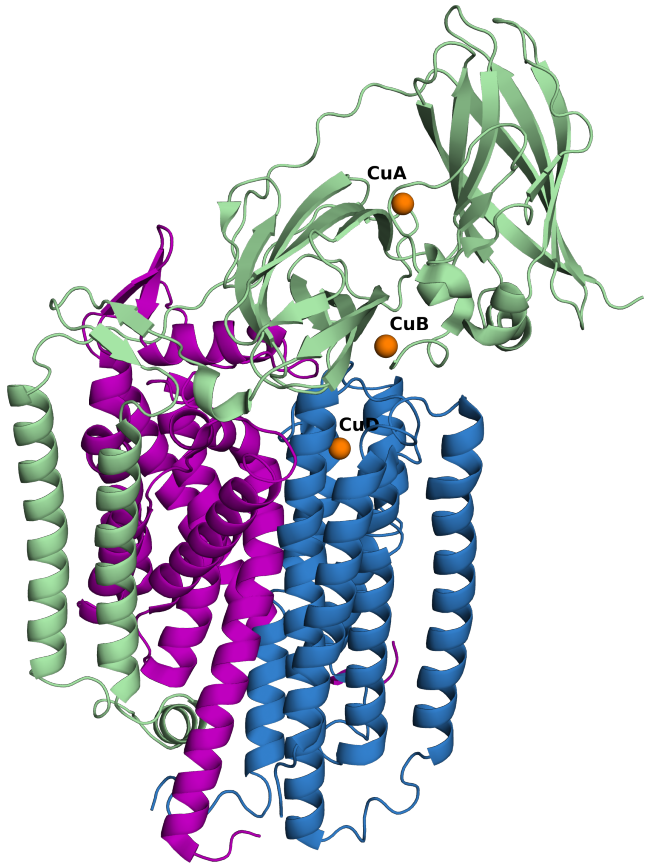



Figure 5.2: A protomer of pMMO from *Methylococcus capsulatus* (Bath) (PDB ID 7S4H) at 2.14 Å resolution, obtained via cryo-EM (Koo et al., 2022). Lipids and water molecules have been omitted for clarity. The three subunits PmoA, PmoB and PmoC are shown in purple, green and blue, respectively. The metal centers $Cu_A$, $Cu_B$ and $Cu_D$ are shown as orange spheres. Figure made with PyMOL (Schrödinger, 2024).

Already in 1994, Chan and coworkers suggested, based on electron paramagnetic resonance (EPR) studies, that the active site of pMMO contains a trinuclear Cu(II) cluster (Nguyen et al., 1994). They found 12–15 Cu ions per protomer and later argued that Rosenzweig's crystallisation protocol leads to a loss of ∼12 copper ions per trimer, indicating that the original crystal structure may not represent the fully copper-loaded state of pMMO (Chan et al., 2007). The same group has also modelled a putative trinuclear active copper site, $Cu_E$, in the crystal structure (Wang et al., 2017; Chang et al., 2021; Chan et al., 2021, 2022).

While $Cu_B$ was originally suggested to be the active site of pMMO (Lieberman and Rosenzweig, 2005a), this has been questioned in later studies (Ross and Rosenzweig, 2017; Ro et al., 2019). A recent QM/MM study suggested that $Cu_B$ may instead be used to produce

$H_2O_2$ (Lundgren et al., 2025). It has also been shown that if the $Cu_C$ site is replaced with a zinc ion, this abolishes the activity of pMMO, indicating that $Cu_C$ is important for the enzyme's function (Sirajuddin et al., 2014). While obviously a considerable amount of effort has been put into finding the active site of pMMO, it still remains controversial.

Recently, the first cryo-EM structures of pMMO were published, at 2.6 Å (Chan and coworkers) and at 2.14–2.46 Å (Rosenzweig and coworkers) (Chang et al., 2021; Koo et al., 2022). With pMMO being a membrane protein, cryo-EM should be a more suitable structure determination method than crystallography, as cryo-EM offers the possibility to better reproduce the native membrane environment. Interestingly, one of Rosenzweig's cryo-EM structures showed a previously unseen copper site, $Cu_D$, in the PmoC subunit, close to the $Cu_C$ site (Koo et al., 2022). Later QM studies of the $Cu_D$ site have suggested that this is the active site of pMMO (Peng et al., 2023).

In this thesis, Chan's structure and two of Rosenzweig's cryo-EM structures have been studied in **Papers II** and **III**.

## 5.4 Ribonucleotide reductase

Ribonucleotide reductases (RNRs) are enzymes that catalyse the conversion of ribonucleotides to deoxyribonucleotides, which are the building blocks of deoxyribonucleic acid (DNA) (Hofer et al., 2012). Three main classes of RNRs are known and they all rely on the generation of a stable intermediate radical. Additionally, all RNRs are allosterically regulated (Eklund et al., 2001).

RNR class I is a heterodimeric tetramer, composed of a large (R1a) and a small (R2a) subunit. The substrate is bound by R1a, whereas R2a contains a di-metal (di-iron or di-manganese) cofactor that generates a tyrosyl radical. This radical is used to initiate a reduction reaction where the 2'-OH-group of the ribonucleotide is reduced to a hydrogen (Hofer et al., 2012; Cotruvo et al., 2013; Greene et al., 2020). Common to all RNR class I enzymes is that they require oxygen for the generation of the radical, which limits their function to aerobic organisms (Nordlund and Reichard, 2006).

RNR class II are either monomeric or dimeric and relies on cobalamin (vitamin B12) for the generation of a radical. This allows them to function under both aerobic and anaerobic conditions (Nordlund and Reichard, 2006; Lundin et al., 2010; Hofer et al., 2012).

RNR class III instead generate a glycyl radical through the use of a S-adenosylmethionine

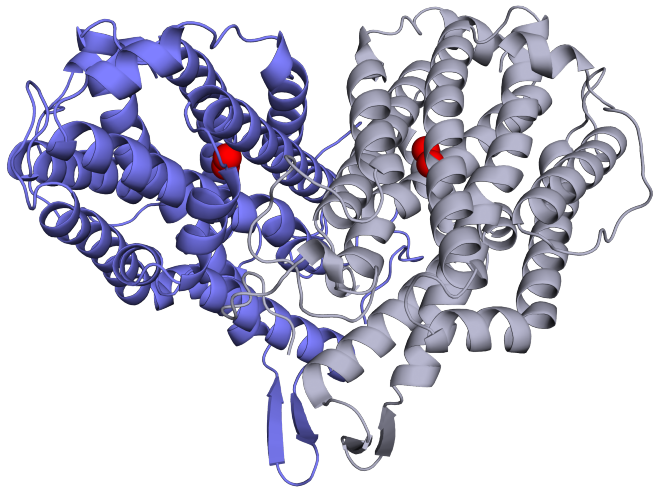


**Figure 5.3:** The dimeric R2a subunit of RNR class I from *Escherichia coli* (PDB ID 1MXR) at 1.42 Å resolution, obtained via XRD (Högbom et al., 2003). Water molecules and crystallisation remnants have been omitted for clarity. The di-iron centers in each subunit are shown as red spheres. Figure made with PyMOL (Schrödinger, 2024).

cofactor and an iron-sulfur cluster. The glycyl radical is sensitive to oxygen and thus RNR class III can only function under anaerobic conditions (Eliasson et al., 1990; Nordlund and Reichard, 2006; Lundin et al., 2010; Hofer et al., 2012).

In this thesis, the di-iron center of the R2a subunit in RNR class I, expressed in *Escherichia coli*, has been studied in **Paper IV**, in both oxidised and reduced forms.

The di-iron cofactor in R2a is coordinated by four carboxylate groups (Asp and Glu) and two histidine residues. In the oxidised form, the di-iron center is in the Fe(III)Fe(III) state and is stabilised by a $\mu$-oxo or a $\mu$-hydroxo bridge between the two Fe ions as well as two water molecules, making the iron ions five or six-coordinate (Nordlund et al., 1990; Nordlund and Eklund, 1993; Högbom et al., 2003). In the reduced form, the di-iron center is in the Fe(II)Fe(II) state. It has been suggested that the bridging ligand is missing in the reduced form, but newer studies have indicated that a bridging solvent molecule is still present, making the irons four or five-coordinate (Logan et al., 1996; Hofer et al., 2025).

Both the oxidised and reduced forms of R2a are stable, where the reduced form in an aerobic environment will slowly be oxidised to the oxidised form. During oxidation, the tyrosyl radical is generated (Nordlund et al., 1990; Logan et al., 1996). The oxidised form can in turn be chemically reduced back to the reduced form (Sahlin et al., 1989).

# Chapter 6

# Summary of papers

## 6.1 Paper I

The crystal structure of Mo-nitrogenase was solved in 1992, while that of V-nitrogenase was solved in 2017 (Kim and Rees, 1992; Einsle, 2014; Sippel and Einsle, 2017). The first crystal and cryo-EM structures of Fe-nitrogenase were published in 2023 (Trncik et al., 2023; Schmidt et al., 2024). These studies have shown that all three nitrogenases involve two proteins, the Fe protein that supplies electrons and the Mo/V/Fe proteins that catalyse the reduction of nitrogen in their Fe/Mo/V/Fe cluster, coordinated by a homocitrate molecule.

In this study, based on an XRD crystal structure (PDB ID 8BOQ) from *Azotobacter vinelandii* at 1.55 Å resolution, the goal was to perform the first QM/MM study of Fe-nitrogenase and to settle the protonation state of the $E_1$ state. A recent EPR study had suggested that the $E_1$ state of Fe-nitrogenase should contain a photolysable hydride ion, in contrast to previous QM/MM studies of Mo-nitrogenase, which have indicated that the $E_1$ state is protonated on a $\mu_2$-sulfide ion (S2B) (Lukoyanov et al., 2022). Setting up a QM/MM study is far from trivial, especially for a complex system such as nitrogenase. In particular, the protonation state of the catalytic FeFe-cluster with all its ligands needs to be settled. This includes the homocitrate ligand, which contains three carboxylate groups and a hydroxyl group. In neutral solution, the carboxylate groups are deprotonated, while the hydroxyl group is protonated. However, in the protein environment, the hydroxyl group binds to one of the Fe ions in the FeFe cluster, which could lower its p$K_a$ value and lead to deprotonation. Estimating p$K_a$ values of chemical groups in proteins with QM or QM/MM methods is quite difficult, as the corresponding study of Mo-nitrogenase showed (Cao et al., 2017). Therefore, we instead performed QR to settle the protonation state of the homocitrate ligand.

An early, local, version of `QRef` was used for the QR calculations, where the FeFe cluster ($Fe_8S_9C$), homocitrate and the coordinating residues Cys-257 (modelled as methyl sulfide, $CSH_3$) and His-423 (modelled as 4-methylimidazole, protonated on N$\varepsilon$2) from the A subunit were included in the QM system. The TPSS functional with def2-SV(P) as the basis set, together with DFT–D4 dispersion correction, was used for the QR calculations. Four different protonation states of the homocitrate ligand were considered, namely protonation on the hydroxyl group (1Ha), protonation on one of the carboxylate groups (1Hc; examination of the surrounding protein indicated that two of the carboxylate groups are most likely deprotonated, whereas the third carboxylate group might be protonated), protonation on both groups (2H) as well as deprotonation of both groups (0H), see figure 6.2. Depending on the protonation state of the homocitrate ligand, the net charge of the QM system was either –7 (0H), –6 (1Ha and 1Hc) or –5 (2H). In order to obtain the proper electronic structure (i.e. anti-ferromagnetically coupled spins, through the use of broken-symmetry (BS) approach), a single-point QM calculation of the system was first run with a multiplicity of 37 (i.e. with all Fe ions in the high-spin state and with all spins aligned) and the desired BS state was then obtained through the use of the `flipspin` procedure in `ORCA` (Noodleman, 1981; Lovell et al., 2001; Neese, 2012). The QR calculations were then started from the converged wave function of these single point calculations, ensuring that the QR calculations remained in the proper BS state. We tested five different weights for the experimental data, $w_{data}$, and found that a value of 1.0 resulted in structures that were significantly affected by both the crystallographic data and the QM calculations.

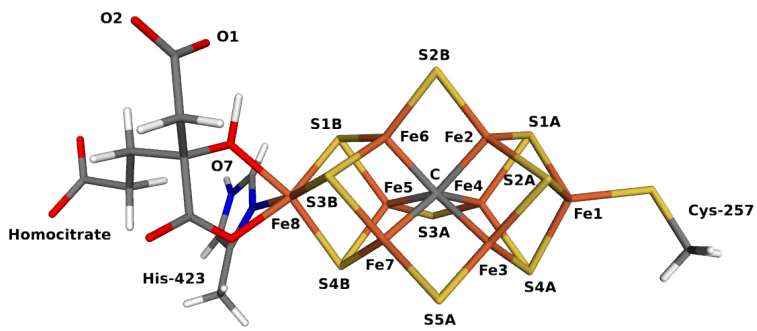



**Figure 6.1:** The FeFe cluster in Fe-nitrogenase, coordinated by Cys-257, His-423 and homocitrate. Figure made with PyMOL (Schrödinger, 2024).

For all four considered protonation states of the homocitrate ligand, BS2358 for the FeFe cluster was used (which is what the pure QM/MM calculations indicated to be the most stable BS state for $E_0$). The four numbers indicate which of the iron atoms have spins pointing in the same direction (corresponding to the labels in figure 6.1). The four missing

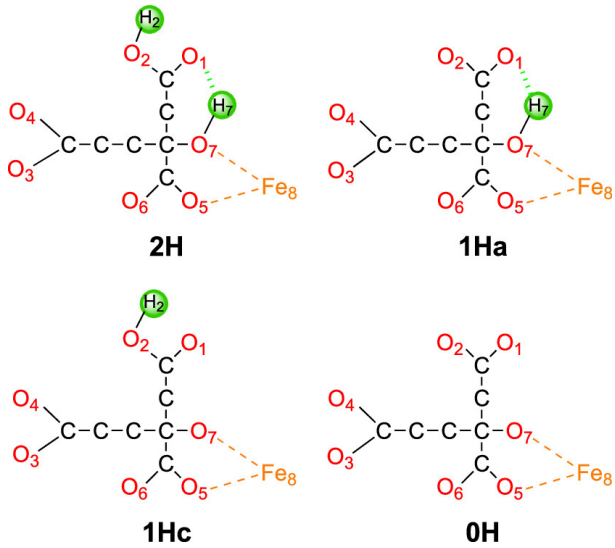



Figure 6.2: The four different protonation states of the homocitrate ligand considered in this study.

Table 6.1: Results of the four QR calculations of Fe-nitrogenase with different protonation states of homocitrate. The quality measures are the real-space Z-scores based on the difference maps (RSZD), the real-space R factors (RSR), the real-space correlation coefficients (RSCC) for homocitrate, the strain energy of the QM region ($\Delta E_{str}$), and the sum of the difference in the Fe–O bond lengths ($\Sigma\Delta d$) to the homocitrate O5 (carboxylate) and O7 (alcohol) atoms in the quantum-refined structure [$d$(Fe–O)$_{QR}$] and in a structure optimised by QM/MM without any crystallographic information [$d$(Fe–O)$_{QM}$]. The best results of each quality measure are marked in bold face.

| | RSZD | RSR | RSCC | $\Delta E_{str}$ kJ/mol | $\Sigma\Delta d$ (Å) | $d$(Fe–O)$_{QR}$ | | $d$(Fe–O)$_{QM}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | O5 | O7 | O5 | O7 |
| 2H | 1.1 | 0.046 | 0.962 | **35** | 0.08 | 2.22 | 2.21 | 2.23 | 2.28 |
| 1Ha | **0.5** | **0.044** | **0.967** | 41 | **0.07** | 2.17 | 2.20 | 2.13 | 2.23 |
| 1Hc | 1.7 | 0.050 | 0.955 | 54 | 0.08 | 2.23 | 2.13 | 2.22 | 2.06 |
| 0H | 2.8 | 0.055 | 0.944 | 38 | 0.10 | 2.22 | 2.06 | 2.19 | 1.99 |

numbers (i.e. 1467 in the BS2358 case) thus have spins pointing in the opposite direction.

The results from the QR calculations are summarised in table 6.1, where all quality measures (except the strain energy, which is not directly comparable between different protonation states) indicate that the 1Ha model (protonation on the hydroxyl group) is the most likely protonation state of the homocitrate ligand. This is in concordance with QM/MM calculations performed in the same study, which showed that the 1Ha model is energetically the most stable protonation state. The results are also in agreement with what was found for Mo-nitrogenase from both QR and QM/MM calculations (Cao et al., 2017).

With this conclusion, the QM/MM calculations could be set up and it was shown that the $E_1$ state of Fe-nitrogenase most likely contains a protonated S2B, in accordance with what was found for Mo-nitrogenase, but in contrast to the findings in the EPR study (Lukoyanov

et al., 2022). These QM/MM calculations were performed by another student in our group.

While `QRef` was still in an early stage of development at the time of this study, the results provided confidence in the method and implementation. The study also highlighted the need for further development (e.g. being able to treat multiple QM systems, as S2B has partial occupancy in 8BOQ), as well as the need to establish best practices for the use of `QRef` (e.g. determining $w_{data}$). These issues were addressed in **Paper II**.

## 6.2 Paper II

**Paper II** is the central work in this thesis. It describes a new implementation of QR, `QRef`, as an interface between the `ORCA` QM software and the `PHENIX` biomolecular structure software (Neese, 2012; Liebschner et al., 2019). Thereby, we solved two issues with the old `ComQum-X/U` interfaces, viz. employing freely available (for academic users) software and using a modern and developing macromolecular structure software (the old `CNS` software is no longer being developed).

Although the final `QRef` interface is rather modest in size (about 650 lines of Python code), it took a considerable amount of effort to develop and test. The switch from `Turbomole` to `ORCA` was trivial and could be done in a few days, owing to the modular structure of both `Turbomole` and `ORCA`. However, switching from `CNS` to `PHENIX` was a major effort due to the black-box nature of the `PHENIX` software, the lack of application programming interface (API) documentation for `cctbx` and minimal support from the `PHENIX` developers.

Compared to our previous QR implementations, `ComQum-X/U`, `QRef` is a complete reimplementation, with additional features. In `QRef`, the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) minimiser in `PHENIX` is used, whereas in `ComQum-X/U` the `Turbomole` BFGS minimiser is used (i.e. in `ComQum-X/U` the minimisation is done from the *model* perspective, in `QRef` the minimisation is done from the *real* perspective). Additionally, in `ComQum-X/U`, hydrogen link atoms are used for the (*low, model*) and (*high, model*) terms, while in `QRef` hydrogen link atoms are used only for the (*high, model*) terms. Chapter 4, sections 4.3.1 and 4.3.3 provide more details about the `QRef` implementation and workflow.

In **Paper I**, an early development version of `QRef` was employed, whereas in **Paper II**, the first released version of `QRef` is described. Compared to **Paper I**, `QRef` was extended to be able to handle an arbitrary number of QM regions. It was also extended to be compatible with real-space refinement, opening up for applications to cryo-EM data sets.

In order to make `QRef` available to the community of structural biologists, we provide the `QRef` interface in a public GitHub repository (`https://github.com/krlun/QRef`). There, we also provide a detailed description of how to use `QRef` for QR calculations, including a number of example applications.

In **Paper II**, we also validated the performance of `QRef` for XRD, ND and cryo-EM and we also started to establish best practices for its use. The test cases also illustrate the benefits of using QR over traditional refinement methods.

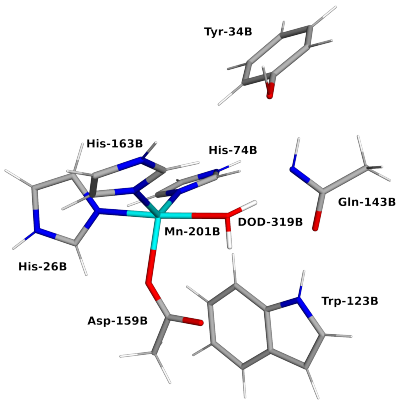
Four test cases were considered:

- A recently released ND structure of reduced MnSOD (PDB ID 7KKW) obtained at 2.30 Å resolution (Azadmanesh et al., 2021). It suggests a fair amount of chemically suspicious features, e.g. a deprotonated glutamine residue.
- An XRD structure of V-nitrogenase (PDB ID 5N6Y) obtained at 1.35 Å resolution (Sippel and Einsle, 2017). It was considered in a previous study, using `ComQum-X`, with the aim to identify an unknown bidentate ligand of the FeV-cluster (Bergmann et al., 2021a). Our aim was to see if `QRef` could reproduce these results and thus be used to discriminate between different interpretations of the structure.
- The XRD structure of Fe-nitrogenase (PDB ID 8BOQ) obtained at 1.55 Å resolution, which was also considered in **Paper I** (Trncik et al., 2023). Again, we studied the protonation state of the homocitrate ligand, but in this study we concentrated on the technical aspects of the calculations, in particular the choice of the $w_{data}$ parameter.
- A cryo-EM structure of pMMO (PDB ID 7S4H, EMDB ID EMD-24826) at 2.14 Å resolution (Koo et al., 2022), for which we focused on the putative $Cu_D$ active site and showed that `QRef` can correct local structural issues (suspicious bond lengths) in cryo-EM structures.

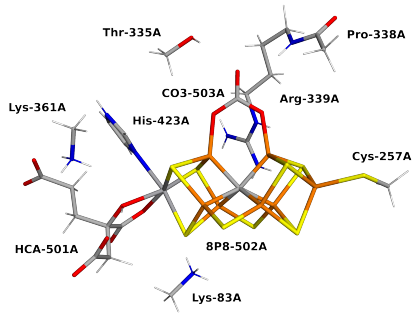
In all test cases, the TPSS functional with the def2-SV(P) basis set was used for all calculations, together with DFT–D4 dispersion correction. For the cryo-EM case, in some of the calculations we also employed a conductor-like polarised continuum model (CPCM) with a dielectric constant of 4.0 in order to mimic the effect of the surrounding protein environment and prevent spurious proton transfers within the QM system (Cammi et al., 2000; Bergmann et al., 2021b). The experimental data sets were chosen from three different sources, XRD, ND and cryo-EM, in order to show that `QRef` works properly and provide results that behave as expected for different types of experimental data.

In the MnSOD case, we showed that that the calculations behave as expected, i.e. that the RSZD scores decrease with increasing $w_{data}$, while the strain energies increase with increasing $w_{data}$. At the same time, the resulting QR geometries come closer to the QM reference geometry with decreasing $w_{data}$, as measured by the sum of the differences in bond lengths ($\Sigma\Delta d$) between the QR structures and the QM reference structure. Additionally, we compared the two different definitions of the strain energy described in section 4.4 and found that they behave similarly. In general, we recommend the use of the definition in equation 4.13 for calculating the strain energy, as it is the most robust and straightforward to use.

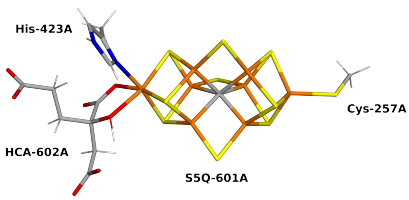
We also found that the deprotonated glutamine in subunit B is in disagreement with



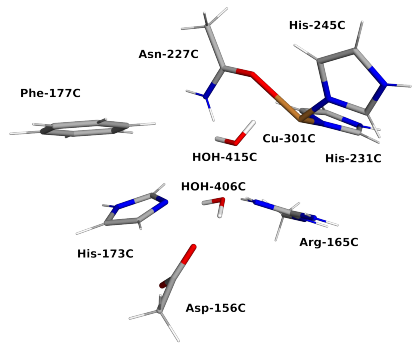

(a) MnSOD, net charge 0, $S = 5$.




(b) V-nitrogenase, net charge –2 or –1 (depending on the unknown ligand), $S = 0$, BS235.




(c) Fe-nitrogenase, net charge –7, –6 or –5, $S = 0$, BS2358.




(d) pMMO, net charge +1, $S = 0$.

**Figure 6.3:** The four different QM regions employed in **Paper II**: (a) MnSOD (PDB ID 7KKW), (b) V-nitrogenase (PDB ID 5N6Y), (c) Fe-nitrogenase (PDB ID 8BOQ) and (d) pMMO (PDB ID 7S4H). Figures made with PyMOL (Schrödinger, 2024).

the QM calculations and for lower values of $w_{data}$ (i.e. when the QM restraints are up-weighted), the glutamine residue becomes protonated during the QR calculations and the solvent water molecule coordinating the Mn ion becomes a hydroxide ion, illustrating that QM calculations are not dependent on a predefined bonding topology, contrary to the case of using force fields in refinement. A thorough assessment of this neutron MnSOD structure, as well as the corresponding oxidised form (PDB ID 7KKS), is the subject of **Paper V**, where the aim is to fully elucidate the protonation states of many more residues and the

nature of the solvent molecules in the active sites.

Furthermore, in order to determine a suitable value for $w_{data}$ when testing different hypotheses regarding the most likely model (the same value of $w_{data}$ should be used for all hypotheses, otherwise the validation metrics between different models are not comparable), we tested a series of QR calculations with different values of $w_{data}$. The conclusion is that a value of $w_{data}$ should be chosen for which the RSZD scores and strain energy are both reasonably low. While this procedure is somewhat subjective, it is still a systematic approach to determine $w_{data}$ and we also note that the results are not overly sensitive to the exact value of $w_{data}$, as long as it is within a reasonable range.

Additionally, we noted that the values suggested for $w_{data}$ by the automatic weight determination procedure in `PHENIX` from a non-QR refinement are typically in the same range as those determined manually, indicating that our suggested value for $w_{QM}$ is adequately calibrated. Based on this, we suggest that the weight from a non-QR refinement can be used as a starting point for $w_{data}$ in QR calculations.

For the V-nitrogenase case, we were able to replicate the results from the previous QR study, again suggesting that out of the three isoelectronic ligands considered, carbonate ($CO_3^{2-}$), bicarbonate ($HCO_3^-$) and nitrate ($NO_3^-$), carbonate is the most likely candidate for the unknown bidentate ligand, based on RSZD scores, as well as RSCC and RSR scores. It should be noted, that even though the starting structure (5N6Y) was of very high quality at 1.35 Å resolution, the authors of the original study were hesitant in identifying the ligand as carbonate, although it was their preferred candidate (Sippel and Einsle, 2017). Additionally, the validation metrics obtained from QR with `QRef` surpassed the previous results, obtained with `ComQum-X`, indicating that `QRef` is a robust and reliable implementation of QR.

In the Fe-nitrogenase case, the results from **Paper I** were reused (no new calculations were performed for this study), but we provided a more thorough discussion of the technical aspects of the calculations, in particular the choice of $w_{data}$. The conclusion from the analysis of Fe-nitrogenase is that a weight where both the experimental quality measures, as well as the strain energies are affected should be used, not only to determine protonation states, but in the general case of using QR.

Finally, for the pMMO case, we showed that `QRef` can be used to improve local structural issues in cryo-EM structures. In the deposited structure (7S4H), His-245C coordinates with N$\varepsilon$2 to Cu at a suspiciously short distance of only 1.50 Å. This distance increases to a more reasonable distance of 1.9–2.0 Å in the QR structures, again highlighting the need for high-quality restraints when using equation 2.42 to calculate $T_{data}$ for low-resolution structures.

Additionally, the $Cu_D$ site contains two water molecules, which are at non-binding distances in the original cryo-EM structure. They showed a tendency to coordinate to the Cu ion in the QR structures, depending on the size of the QM region. One of the water molecules forms hydrogen bonds to the surrounding residues and if these are included in the QM region, this water molecule forms a stable structure without coordinating to the Cu ion. However, the other water molecule does not have any hydrogen-bonding partners and it is then natural that it may bind to the Cu ion. It is also possible that this water molecule may have a large mobility and may occupy several different positions in the cryo-EM structure.

We saw a deterioration of the RSCC scores in all cryo-EM test cases and that the water molecules and the Cu ion moved somewhat out of the density. This is actually expected, as the positions of atoms in a refined structure are a compromise between the experimental data and the restraints, as dictated by equation 2.31. However, for metal ions and water molecules, there are essentially no empirical restraints in a non-QR refinement (besides a repulsive van der Waals term, active only at short distances). Therefore, metals and water molecules are essentially overfitted towards the experimental data in non-QR structures. This was checked by studying how much the waters and the Cu ion move out of the density in the QR structures compared to other atoms in the structure, which showed no significant difference.

The pMMO structure (7S4H) used in this study, as well as two other cryo-EM structures of pMMO (7S4J and 7EV9) are the subject of a more thorough study in **Paper III**.

In conclusion, this validation study showed that the `QRef` interface is a robust and reliable implementation of QR, which can be used for both XRD and ND data in reciprocal space refinement. Moreover, for the first time, QR was successfully applied to a metalloprotein in real-space refinement. We also suggest that QR with a subtractive hybrid scheme is the method of choice to get proper structures of metal sites in cryo-EM data.

## 6.3 Paper III

With the performance of `QRef` validated and the interface having reached a level where it can be used for applications, in this third paper we performed a more thorough study of three recently published cryo-EM structures of pMMO from *Methylococcus capsulatus* (Bath) at 2.6 Å (PDB ID 7EV9, EMDB ID EMD-31325), 2.14 Å (PDB ID 7S4H, EMDB ID EMD-24826) and 2.16 Å (PDB ID 7S4J, EMDB ID EMD-24828), respectively (Chang et al., 2021; Koo et al., 2022).

7EV9 was published in 2021 by Chan and coworkers, whereas the two structures 7S4H and 7S4J (together with six other structures) were published in 2022 by Rosenzweig and coworkers. The $Cu_A$ and $Cu_B$ sites are present in all three deposited structures, with similar geometries as in the crystal structures of pMMO. $Cu_A$ is the least controversial site and in all three structures modelled as mononuclear. In 7S4H and 7S4J, $Cu_B$ is modelled as mononuclear, whereas in 7EV9 it is modelled as dinuclear. In 7S4J, $Cu_C$ is present, whereas it is absent in 7S4H. In 7S4H, the opposite holds true, i.e. $Cu_D$ is present and $Cu_C$ is absent. $Cu_D$ is a putative new metal center and is close to $Cu_C$ ($\sim$ 6 Å away) and seems to require the stabilising effect of lipids to form. In 7EV9, neither $Cu_C$ nor $Cu_D$ is present.

In 7EV9, three additional copper sites are modelled in the deposited structure, $Cu_E$, Cu504 and Cu505/506. $Cu_E$ is the active site of pMMO proposed by Chan and coworkers in the PmoA subunit, with the suggestion of a trinuclear copper cluster. Until the publication of 7EV9, $Cu_E$ had not been observed in any structure. In 7EV9 it is modelled as a dinuclear copper cluster, with the third copper ion missing. Cu504 is part of the so-called "copper-sponge" and modelled as mononuclear. Cu505/506 is also proposed to be part of the copper-sponge and is modelled as a dinuclear copper site.

For all metal sites in all models, we performed QR calculations with `QRef`, using the TPSS functional with a def2-SV(P) basis set and DFT–D4 dispersion correction. In some cases we also employed CPCM with a dielectric constant of 4.0, in order to prevent spurious proton transfers within the QM system. The results were evaluated through RSCC scores for each residue as calculated by `phenix.map_model_cc`, strain energies, changes in key distances, visual inspection of the resulting models compared to the electrostatic densities, as well as through the use of the `CheckMyMetal` web server (Zheng et al., 2017a; Afonine et al., 2018a).

The setup of the QR jobs followed the procedure outlined in chapter 4, section 4.3.3. The QM regions were chosen to include the metal ion(s) and at least all first-sphere coordinating residues, as well as any solvent water molecules near a copper ion.

The oxidation states of the Cu ions are not reported in the original publications. Therefore, we tested either Cu(I) and Cu(II) for all copper ions. In the case of dinuclear copper sites, all permutations of these oxidation states were tested (i.e. Cu(I)Cu(I), Cu(I)Cu(II) and Cu(II)Cu(II)). For copper sites that were flagged by `CheckMyMetal` to be dubious ($Cu_E$, Cu504 and Cu505/506), we also tested replacing one or both of the copper ions with water molecules.

The best interpretations from our QR calculations are summarised in table 6.2, whereas the full results are provided in **Paper III**. For all structures, $Cu_A$ was found to be a mononuclear Cu(I) site with two (7EV9) or three ligands (7S4H and 7S4J). QR confirmed that $Cu_B$ is a mononuclear site in the 7S4H and 7S4J structures and indicated this is the case also in the 7EV9 structure, although it was modelled as dinuclear in the deposited structure. $Cu_C$ was confirmed to be a mononuclear Cu(I) site in 7S4J. Likewise, $Cu_D$ is mononuclear in 7S4H. However, as discussed also in **Paper II**, there is a water molecule that might coordinate at least intermittently to $Cu_D$, depending on the oxidation state of the Cu ion and dynamic effects. On the other hand, we found no support for any Cu ions in the remaining three sites in the 7EV9 structure. Instead, they were much better modelled by one or two water molecules.

In conclusion, this study shows that QR can improve the modelling of Cu ions in cryo-EM structures and it can be used to discriminate between correctly and incorrectly assigned metal ions. Therefore, we suggest that QR should be the method of choice to model metal sites in cryo-EM structures, solving the problem that essentially no empirical restraints are used for metals in standard refinement.

**Table 6.2:** The most likely interpretations of the copper sites in the three cryo-EM structures of pMMO, as suggested by our QR calculations.

| | $Cu_A$ | $Cu_B$ | $Cu_C$ | $Cu_D$ | $Cu_E$ | Cu504 | Cu505/506 |
|---|---|---|---|---|---|---|---|
| 7EV9 | Cu(I) | Cu(II) | n/a | n/a | $2H_2O$ | $H_2O$ | $2H_2O$ |
| 7S4H | Cu(I) | Cu(II) | n/a | Cu(I) or Cu(II) | n/a | n/a | n/a |
| 7S4J | Cu(I) | Cu(II) | Cu(I) | n/a | n/a | n/a | n/a |

## 6.4 Paper IV

In this study, we for the first time used QR for data originating from X-ray free-electron laser (XFEL) and microcrystal electron diffraction (MicroED) experiments. While XFEL data is expected to behave in the same way as conventional XRD data, the implementation of `QRef` in `PHENIX` made it possible to extend QR to ED. However, a caveat is that only neutral scattering factors for electrons are implemented in `PHENIX`, which may affect the results for charged species.

As a test case, we used six different structures of the RNR class I R2a protein and concentrated on the di-iron site in either the oxidised Fe(III)Fe(III) or the reduced Fe(II)Fe(II) state. For each oxidation state, we used three different structures, obtained from conventional XRD experiments (already published, 1MXR (oxidised, 1.42 Å) and 1XIK (reduced, 1.70 Å)), XFEL (9SIG (oxidised, 1.9 Å) and 9SIH (reduced, 1.7 Å)) and MicroED (oxidised, 2.00 Å and reduced, 2.20 Å), respectively. The XFEL and MicroED data were kindly provided by our collaborators at Stockholm University (Pacoste, 2025).

With six data sets in total (three for each oxidation state) for the same protein, we could compare the results from QR between different experimental techniques and see if they perform similarly. For each structure we tested different protonation states of the iron-bridging solvent-derived ligand, in order to determine its protonation state.

For each of the six structures, the two iron ions and the first-sphere ligands were included in the QM region (Asp-84, Glu-115, His-118, Glu-204, Glu-238 and His-241), as well as the bridging solvent-derived ligand (except for the 1XIK structure, where it is absent), modelled as either $O^{2-}$, $OH^{-}$ or $H_2O$. In the case of the oxidised structures, the two iron coordinating water molecules were also included in the QM region (they are missing in the reduced structures). For 1MXR, 1XIK and 9SIG we used subunit A, for 9SIH subunit B (as this was better resolved) and for the MicroED structures subunit C. The TPSS functional with a def2-SV(P) basis set and DFT–D4 dispersion correction was used for all calculations. The iron ions were assumed to be in a low-spin state ($S = 0$), with antiferromagnetic coupling between the two iron ions. This was achieved, like in **Paper I** and **Paper II**, by using the BS approach through the `flipspin` procedure in `ORCA` (Noodleman, 1981; Lovell et al., 2001; Neese, 2012).

The results were evaluated through RSZD scores for each residue as calculated by `EDSTATS` (Tickle, 2012), strain energies, changes in Fe–ligand distances, changes in the coordination of the ligating residues to the iron centers, as well as visual inspection of the resulting models and the corresponding difference density maps.

**Table 6.3:** The most likely protonation states of the bridging ligand in the six different RNR structures, modelled as oxide ion ($O^{2-}$), hydroxide ion ($OH^-$) or water molecule ($H_2O$), according to our QR calculations.

| Oxidised (Fe(III)Fe(III)) | | | Reduced (Fe(II)Fe(II)) | | |
|---|---|---|---|---|---|
| 1MXR | 9SIG | MicroED | 1XIK | 9SIH | MicroED |
| $O^{2-}$ or $OH^-$ | $O^{2-}$ | $OH^-$ | n/a | $H_2O$ | $H_2O$ |

The results for the protonation states of the bridging ligand are summarised in table 6.3.

While our results point towards the possibility of the bridging ligand being either an oxide or a hydroxide ion in the oxidised state of R2a for 1MXR and a hydroxide ion in the MicroED structure, they strongly suggest it is an oxide ion in the XFEL 9SIG structure. Previous spectroscopic studies have suggested that the bridging ligand is an oxide ion in the oxidised state of R2a (Sjöberg et al., 1982; Scarrow et al., 1986; Bunker et al., 1987). We thus interpret the results for 1MXR and the MicroED structures as indicative that the metal site has been photo-reduced during data collection and that XFEL data collection, with its ultrashort pulses, indeed operates in a diffraction-before-destruction regime.

For the reduced state of R2a, our results clearly suggest that the bridging ligand is a water molecule in both 9SIH and the MicroED structure. Additionally, in the 1XIK structure, there are several suspiciously short Fe–O distances of about 1.7 Å in the deposited structure. QR corrects these distances to about 2.0 Å.

In conclusion, this study shows that QR can be successfully applied to ED as well as XFEL data. While the expectation was that the XFEL data would behave similarly to data obtained from XRD, it was not obvious that QR would work for ED data, as there is a higher charge dependence on the scattering factors for electrons compared to X-rays. However, our results indicate that QR works well for ED data and that it can be used to improve the modelling of metal sites in ED structures as well.

## 6.5 Paper V

In this final paper of the thesis, we performed a thorough study of two perdeuterated ND structures of oxidised (Mn(III), PDB ID 7KKS, 2.20 Å) and reduced (Mn(II), PDB ID 7KKW, 2.30 Å) human MnSOD, published in 2021 by Borgstahl and coworkers (Azadmanesh et al., 2021). The deposited structures show a number of unusual chemical features, e.g. deprotonated[1] glutamine, histidine and tyrosine residues. Additionally, there are several notable differences between the two subunits of the dimeric enzyme, e.g. the presence of a second $OH^-$ ligand in the active site of one subunit but not the other, in the reduced structure.

With the scattering length of deuterium being on the same order of magnitude as those of other common elements in proteins (e.g. C, N and O), ND is a powerful technique to determine protonation states and hydrogen-bonding networks. However, ND data sets are often of lower resolution than XRD data sets, due to the low flux of neutron sources and in turn longer data collection times, as well as geometrically limiting instrumental setups that reduce the usable high-angle signal. Additionally, the introduction of hydrogens in the structure increases the number of parameters to refine substantially, further complicating the refinement process. Therefore, high-quality restraints are essential to obtain chemically sensible structures from ND data. QR is thus an ideal method to use when refining ND structures.

In all cases, like in the previous papers, the TPSS functional with a def2-SV(P) basis set and DFT–D4 dispersion correction was used for all calculations. When manganese was included in the QM region, it was assumed to be in the high-spin state ($S = 2$ for Mn(III) and $S = 5/2$ for Mn(II)).

The aim of this study was twofold, viz. to determine the nature and number of solvent molecules in the active site of subunit A in the reduced structure, as well as to elucidate the protonation state of a hydrogen-bonding network (consisting of His-30, Tyr-34, Gln-143, Tyr-166, as well as one or two solvent molecules) nearby the active site, for both subunits, in the two structures. As His-30 interacts with a symmetry-related residue, Tyr-166, from the other dimer in the crystal, `QRef` was extended to be able to handle symmetry interactions for this study.

For each of the two structures, we first performed a scan of $w_{data}$ values, in order to determine a suitable value for this parameter. In both structures this was done for the manganese active site in subunit B, with residues His-26, Tyr-34, His-74, Trp-123, Gln-143, Asp-159

[1] While technically incorrect use of this word in this context, as the structures are perdeuterated, in the following I will for simplicity talk about "protons" and "hydrogens", or variations thereof.

and His-163 included in the QM region, as well as the manganese ion and a coordinating solvent molecule ($OH^-$ in 7KKS and $H_2O$ in 7KKW). The results indicated that $w_{data} = 3$ was a suitable value for both structures and this value was then used for all subsequent QR calculations. A similar value of $w_{data}$ was obtained from a non-QR refinement and we now feel confident in forfeiting the $w_{data}$ scan in future applications, instead relying on the weight from a non-QR refinement as a starting point.

We then performed a series of QR calculations[2]. For 7KKS, we examined the hydrogen-bonding network and in total tested 24 different protonation states, rotations of water molecules and restrained models in subunit A. The corresponding number of different models for subunit B was 20.

For 7KKW subunit B, we instead built a large QM region, consisting of the active site as well as the hydrogen-bonding network, because we were interested also in the protonation state of the Mn-bound solvent molecule and Gln-143. This resulted in 31 different models. Turning our attention to subunit A and wise from the experiences from subunit B, we tried to shortcut the process and instead opted to run two separate sets of calculations, one for the active site (with three different interpretations of the solvent molecules, $2OH^-$, $OH^-$ and $H_2O$) and one set for the hydrogen-bonding network (with 32 different models). The results for the solvent molecule were unfortunately inconclusive, possibly owing to a partial occupancy of the second solvent molecule. Therefore, the best five models from the calculations of the hydrogen-bonding network were tested for different interpretations of the solvent molecules of the active site, resulting in 15 models for this combined QM region, consisting of ∼100 atoms.

The results were evaluated through RSZD scores for each residue as calculated by `EDSTATS` (Tickle, 2012), strain energies, changes in key distances, as well as visual inspection of the resulting models and the corresponding difference density maps.

The metrics of our models show improvement in almost all aspects for all of our tested models compared to the deposited structures, indicating that QR is a powerful method for refining ND structures. Our best interpretations together with the suggestion in the deposited structures are summarised in table 6.4 and the full results are provided in **Paper V**. It should be noted that owing to the low resolution of the neutron data and the problem that strain energies are not directly comparable for structures with different net charge, the various structural interpretations often give a similar result and differences in the RSZD scores were minimal. Therefore, we cannot claim that our suggestions are significantly better than the original structures. However, our results show that more chemically reasonable

[2]The number of different tested models listed in this summary represents the actual number of models I tested; some of these converged to the same structure. For this reason, the number of models presented in the paper itself are in some cases slightly lower.

**Table 6.4:** Our best interpretation of the protonation states of the hydrogen-bonding network in the two subunits of oxidised and reduced MnSOD, as well as the nature of the solvent molecule(s) in the active site of reduced MnSOD (solv), as suggested by our QR calculations. The notations for the protonation states are as follows: His-30: H$\delta$1: proton on N$\delta$1. Tyr-34 and Tyr-166: H$\eta$: proton on phenolic O, the number that follows indicates to which side in the phenol plane the proton points (1 = same side as C$\varepsilon$1, 2 = same side as C$\varepsilon$2), oop indicates that the proton is out of the phenolic plane. Gln-143: H$\varepsilon$2: proton on N$\varepsilon$2, the number (1 or 2) that follows indicates to which side in the amide plane the protons point. The suggestion in the deposited structure is shown in brackets (if it differs from the QR suggestions). * indicates that the proton is missing in the deposited structure, but that it should be there according to the original article (Azadmanesh et al., 2021).

| | Oxidised (Mn(III)) | | Reduced (Mn(II)) | |
|---|---|---|---|---|
| | Subunit A | Subunit B | Subunit A | Subunit B |
| His-30 | H$\delta$1 | H$\delta$1 (no) | H$\delta$1 | H$\delta$1 |
| Tyr-34 | H$\eta$1 (no) | H$\eta$2 (no) | H$\eta$1 (no) | H$\eta$1 (oop) |
| Gln-143 | H$\varepsilon$21 & H$\varepsilon$22 | H$\varepsilon$21 & H$\varepsilon$22 | H$\varepsilon$21 & H$\varepsilon$22 | H$\varepsilon$21 & H$\varepsilon$22 (H$\varepsilon$22) |
| Tyr-166 | H$\eta$1 (no*) | H$\eta$1 (no*) | H$\eta$1 | H$\eta$1 (no) |
| solv | | | $H_2O$ ($2OH^-$) | $H_2O$ |

interpretations are at least as likely as the original protonation states, showing that there is no strong experimental evidence for the questionable chemical features in the deposited structures. In particular, we point out that the fact that if a proton is not observed in a neutron structure, this does not prove that it is not there. Instead, it may indicate that it has several positions or is mobile.

In conclusion, this study shows that QR can be used to improve the modelling of metal sites and hydrogen-bonding networks in neutron structures. However, even with QR it is still challenging to unequivocally determine protonation states in complicated hydrogen-bonding networks, especially when the experimental data is of limited resolution. Additional complementary techniques may be needed to fully resolve such systems.

# Chapter 7

# Conclusions and Outlook

The goals for this thesis set out in chapter 1 have largely been achieved. As a result, a new QR interface, `QRef`, that uses modern macromolecular refinement software (`PHENIX`), is now available to the structural biology community. Compared to some other recent QR implementations, `QRef` also makes it possible to use QR, in a subtractive hybrid scheme, for (complex) metal sites. While this is also possible in the `ComQum-X/U` series of interfaces, `ComQum-X/U` is limited to XRD and ND data in reciprocal space refinement. `QRef` extends reciprocal space QR to also cover ED data, as well as extends QR to real-space refinement for cryo-EM data.

`QRef` is distributed freely and I have had the FAIR principles in mind throughout the development, ensuring that it is findable, accessible, interoperable and reusable (Wilkinson et al., 2016). Additionally, with the license chosen for `QRef` (BSD-3-Clause), users are free to use, modify, build upon and distribute their derivative code as they see fit, as long as proper attribution is given.

It can perhaps be argued that we are overcomplicating the refinement process of macromolecules by introducing QR and subjecting the user to the intricacies of QM calculations. However, from a Bayesian perspective, QR is introducing a more accurate prior (i.e. the QM calculations) into the refinement process. With limited sampling of the likelihood (i.e. the experimental data), it is essential to have a high-quality prior as it will heavily influence the ability to obtain a good posterior (i.e. the refined structure). As a consequence of this, in **Paper I** we successfully used `QRef` to determine the protonation state of a homocitrate ligand in a crystal structure of a very complex enzyme. In paper **Paper II**, we thoroughly validated the performance of `QRef` and also showed that QR can improve the modelling of metal sites in cryo-EM structures, where essentially no empirical restraints are used for metals in standard refinement. In **Paper III**, we performed a more thorough study of three

cryo-EM structures of pMMO, showing that QR can be used to discriminate between correctly and incorrectly assigned metal ions. In **Paper IV**, we applied `QRef` to data from XFEL and ED experiments, showing that QR can be used to improve the modelling of metal sites in ED structures as well, even though ED data typically show more signs of photoreduction (which we were also able to detect from the QR analysis), suffers from multiple scattering effects due to the strong interaction of electrons with matter, as well as that only atomic form factors for neutral species are typically available for ED (Saha et al., 2022). Finally, in **Paper V**, we applied `QRef` to two ND structures of MnSOD that contains a fair amount of chemically suspicious features. Unfortunately, the experimental data was not accurate enough to allow us to with full confidence elucidate the protonation states in the hydrogen-bonding network in these structures. However, we showed that chemically reasonable structures are at least as likely as the original structures.

Additionally, while it has not been used in this thesis, `QRef` supports multiple QM regions, which can be overlapping or disjunct. This allows for QR treatment of time-resolved data sets, a line of research which is currently being pursued by our research group and is already showing very promising results.

As for extensions and further development of QR, the most obvious step seems to be to include electrostatic embedding in `QRef`. This would allow for more accurate treatment of charged species and polar environments, which are common in biological systems. However, this is a challenging task, as it requires obtaining partial charges for the region not treated with QM. Especially, the model of the protein needs to be protonated correctly, which is not always straightforward (because protons are seldom discerned in XRD structures). Another route to go down, which is perhaps not QR in itself but rather more akin to the standard refinement protocol, would be to obtain more accurate force fields (where not just the ideal values are used, but also calibrated force constants) for metals and other challenging moieties, tailored for their specific environments. This can be done through for example the Seminario or the more recent Joyce methods, which derive force field parameters from QM calculations (Seminario, 1996; Vilhena et al., 2021; Giannini et al., 2025). These parameters could then be used in standard refinement, without the need to perform QM calculations *in situ* during refinement. I am personally particularly interested in exploring the Seminario method. Another venue is of course using machine learning potentials, which unfortunately still struggle with metals, but are rapidly improving (Behler, 2016; Mueller et al., 2020; Unke et al., 2021; Chmiela et al., 2023; Novelli et al., 2025).

Furthermore, with the electron density in a sense obtained "for free" in QR calculations, it seems wasteful to not use it, as is currently the case for QR in general. It would be interesting to use the electron density to create tailor-made atomic form factors for the specific chemical environment of each atom in the QM region through Hirshfeld partitioning and leave the IAM behind, as can routinely be done in small-molecule crystallography

(Hirshfeld, 1977; Capelli et al., 2014; Kleemiss et al., 2020). This would likely improve the modelling of the electron density further and could potentially be used also outside of QR, i.e. in standard refinement. Implementations of Hirshfeld atom refinement (HAR), even a fragmentation version (fragHAR), for macromolecular crystallography does already exist (Bergmann et al., 2020).

Another way to construct aspherical atomic form factors is the multipole formalism, which has been attempted in macromolecular crystallography (Hansen and Coppens, 1978; Afonine et al., 2007). Compared to multipole formalism, HAR risks no overfitting, as no additional parameters are introduced, making it more suitable for the typically lower-resolution macromolecular data.

Unfortunately, neither HAR nor multipole formalism has yet seen widespread use in macromolecular crystallography, likely owing to the increased complexity and computational cost, and in particular lack of implementations in macromolecular mainstream software packages.

Finally, `QRef` is not perfect. I tried making it rather user-friendly, but it is command-line based and requires some knowledge of both macromolecular refinement and QM calculations for effective use. It is also an *ad hoc* implementation, that requires some modification of the code in `PHENIX`, where the licensing of `PHENIX` is a bit restrictive. For these reasons, I have started developing a successor to `QRef`, tentatively called `QRef2`, implemented as a `Jupyter` notebook. `QRef2` will be leveraging fully open-source libraries with permissive licenses, such as `cctbx` (which is the library that `PHENIX` is built upon) and/or `GEMMI` for the macromolecular part (Grosse-Kunstleve et al., 2002; Grosse-Kunstleve and Adams, 2002; Wojdyr, 2022), the Atomic Simulation Environment (ASE) (Hjorth Larsen et al., 2017), which allows for efficient communication with a variety of QM software packages, and `SciPy` (Virtanen et al., 2020) and/or `NLopt` (Johnson, 2007) for the optimisation routines.

With `QRef2` implemented as a `Jupyter` notebook, it should be possible to efficiently guide the user through the setup of QR calculations in an interactive manner, with explanations and visualisations along the way. Additionally, through the use of ASE, it will be possible to easily switch between different QM software packages, depending on the user's preferences and available licenses.

I thus envision `QRef2` to be a more user-friendly and flexible QR interface. This would hopefully lower the barrier of entry for QR and make it more accessible to a wider range of structural biologists.

# References


I. A. Abreu and D. E. Cabelli. Superoxide dismutases-a review of the metal-associated mechanistic variations. *Biochim. Biophys. Acta*, 1804(2):263–274, Feb. 2010. doi: 10.1016/j.bbapap.2009.11.005.

M. R. Aczel. What is the nitrogen cycle and why is it key to life? *Front. Young Minds*, 7, Mar. 2019. doi: 10.3389/frym.2019.00041.

P. D. Adams, N. S. Pannu, R. J. Read, and A. T. Brünger. Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc. Natl. Acad. Sci. U. S. A.*, 94(10):5018–5023, May 1997. doi: 10.1073/pnas.94.10.5018.

P. D. Adams, R. W. Grosse-Kunstleve, L.-W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, and T. C. Terwilliger. *PHENIX*: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Struct. Biol.*, 58(11):1948–1954, Nov. 2002. doi: 10.1107/s0907444902016657.

P. V. Afonine, R. W. Grosse-Kunstleve, and P. D. Adams. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr. D Struct. Biol.*, 61(Pt 7):850–855, July 2005. doi: 10.1107/S0907444905007894.

P. V. Afonine, R. W. Grosse-Kunstleve, P. D. Adams, V. Y. Lunin, and A. Urzhumtsev. On macromolecular refinement at subatomic resolution with interatomic scatterers. *Acta Crystallogr. D Biol. Crystallogr.*, 63(Pt 11):1194–1197, Nov. 2007. doi: 10.1107/S0907444907046148.

P. V. Afonine, N. Echols, R. W. Grosse-Kunstleve, N. W. Moriarty, and P. D. Adams. Improved weight optimization in *phenix.refine*. *Comput. Crystallogr. Newsl.*, 2(2):99–103, 2011.

P. V. Afonine, R. W. Grosse-Kunstleve, P. D. Adams, and A. Urzhumtsev. Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Crystallogr. D Struct. Biol.*, 69(Pt 4):625–634, Apr. 2013. doi: 10.1107/S0907444913000462.

P. V. Afonine, B. P. Klaholz, N. W. Moriarty, B. K. Poon, O. V. Sobolev, T. C. Terwilliger, P. D. Adams, and A. Urzhumtsev. New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr. D Struct. Biol.*, 74(Pt 9):814–840, Sept. 2018a. doi: 10.1107/S2059798318009324.

P. V. Afonine, B. K. Poon, R. J. Read, O. V. Sobolev, T. C. Terwilliger, A. Urzhumtsev, and P. D. Adams. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D Struct. Biol.*, 74(Pt 6):531–544, June 2018b. doi: 10.1107/S2059798318006551.

P. V. Afonine, P. D. Adams, and A. G. Urzhumtsev. Efficient structure-factor modeling for crystals with multiple components. *Acta Crystallogr. A Found. Adv.*, 79(Pt 4):345–352, July 2023. doi: 10.1107/S205327332300356X.

D. S. Albers and M. Flint Beal. Mitochondrial dysfunction and oxidative stress in aging and neurodegenerative disease. In *Advances in Dementia Research*, pages 133–154. Springer Vienna, Vienna, 2000. doi: 10.1007/978-3-7091-6781-6_16.

E. Arnold and M. G. Rossmann. The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. *Acta Crystallogr. A*, 44(3):270–283, May 1988. doi: 10.1107/S0108767387011875.

P. W. Atkins and R. S. Friedman. *Molecular Quantum Mechanics*. Oxford University Press, London, England, 5 edition, Nov. 2010. ISBN 9780199541423.

J. Azadmanesh and G. Borgstahl. A review of the catalytic mechanism of human manganese superoxide dismutase. *Antioxidants (Basel)*, 7(2):25, Jan. 2018. doi: 10.3390/antiox7020025.

J. Azadmanesh, W. E. Lutz, L. Coates, K. L. Weiss, and G. E. O. Borgstahl. Direct detection of coupled proton and electron transfers in human manganese superoxide dismutase. *Nat. Commun.*, 12(1):2079, Apr. 2021. doi: 10.1038/s41467-021-22290-1.

J. Azadmanesh, K. Slobodnik, L. R. Struble, W. E. Lutz, L. Coates, K. L. Weiss, D. A. A. Myles, T. Kroll, and G. E. O. Borgstahl. Revealing the atomic and electronic mechanism of human manganese superoxide dismutase product inhibition. *Nat. Commun.*, 15(1):5973, July 2024. doi: 10.1038/s41467-024-50260-w.

K. H. Babai, F. Long, M. Malý, K. Yamashita, and G. N. Murshudov. Improving macromolecular structure refinement with metal-coordination restraints. *Acta Crystallogr. D Struct. Biol.*, 80(Pt 12):821–833, Dec. 2024. doi: 10.1107/S2059798324011458.

T. S. Baker and R. Henderson. Electron cryomicroscopy of biological macromolecules. In *International Tables for Crystallography*, pages 593–614. International Union of Crystallography, Chester, England, Apr. 2012. doi: 10.1107/97809553602060000872.

I. Barbarin-Bocahu and M. Graille. The x-ray crystallography phase problem solved thanks to AlphaFold and RoseTTAFold models: a case-study report. *Acta Crystallogr. D Struct. Biol.*, 78(Pt 4):517–531, Apr. 2022. doi: 10.1107/S2059798322002157.

A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A Gen. Phys.*, 38(6):3098–3100, Sept. 1988. doi: 10.1103/physreva.38.3098.

J. Behler. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.*, 145(17):170901, Nov. 2016. doi: 10.1063/1.4966192.

B. Benediktsson and R. Bjornsson. Analysis of the geometric and electronic structure of spin-coupled iron-sulfur dimers with broken-symmetry DFT: Implications for FeMoco. *J. Chem. Theory Comput.*, 18(3):1437–1457, Mar. 2022. doi: 10.1021/acs.jctc.1c00753.

J. Bergmann, M. Davidson, E. Oksanen, U. Ryde, and D. Jayatilaka. fragHAR: towards ab initio quantum-crystallographic x-ray structure refinement for polypeptides and proteins. *IUCrJ*, 7(Pt 2):158–165, Mar. 2020. doi: 10.1107/S2052252519015975.

J. Bergmann, E. Oksanen, and U. Ryde. Quantum-refinement studies of the bidentate ligand of v-nitrogenase and the protonation state of CO-inhibited mo-nitrogenase. *J. Inorg. Biochem.*, 219(111426):111426, June 2021a. doi: 10.1016/j.jinorgbio.2021.111426.

J. Bergmann, E. Oksanen, and U. Ryde. Can the results of quantum refinement be improved with a continuum-solvation model? *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.*, 77(6):906–918, Dec. 2021b. doi: 10.1107/S2052520621009574.

J. Bergmann, E. Oksanen, and U. Ryde. Combining crystallography with quantum mechanics. *Curr. Opin. Struct. Biol.*, 72:18–26, Feb. 2022. doi: 10.1016/j.sbi.2021.07.002.

H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nat. Struct. Biol.*, 10(12):980, Dec. 2003. doi: 10.1038/nsb1203-980.

H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, 35 (Database issue):D301–3, Jan. 2007. doi: 10.1093/nar/gkl971.

H. Bethe. Zur theorie des durchgangs schneller korpuskularstrahlen durch materie. *Ann. Phys.*, 397(3):325–400, Jan. 1930. doi: 10.1002/andp.19303970303.

M. P. Blakeley, P. Langan, N. Niimura, and A. Podjarny. Neutron crystallography: opportunities, challenges, and limitations. *Curr. Opin. Struct. Biol.*, 18(5):593–600, Oct. 2008. doi: 10.1016/j.sbi.2008.06.009.

O. Borbulevych, R. I. Martin, and L. M. Westerhoff. High-throughput quantum-mechanics/molecular-mechanics (ONIOM) macromolecular crystallographic refinement with PHENIX/DivCon: the impact of mixed hamiltonian methods on ligand and protein structure. *Acta Crystallogr. D Struct. Biol.*, 74(Pt 11):1063–1077, Nov. 2018. doi: 10.1107/S2059798318012913.

O. Y. Borbulevych, J. A. Plumley, R. I. Martin, K. M. Merz, Jr, and L. M. Westerhoff. Accurate macromolecular crystallographic refinement: incorporation of the linear scaling, semiempirical quantum-mechanics program DivCon into the PHENIX refinement package. *Acta Crystallogr. D Struct. Biol.*, 70(Pt 5):1233–1247, May 2014. doi: 10.1107/S1399004714002260.

G. E. Borgstahl, H. E. Parge, M. J. Hickey, W. F. Beyer, Jr, R. A. Hallewell, and J. A. Tainer. The structure of human mitochondrial manganese superoxide dismutase reveals a novel tetrameric interface of two 4-helix bundles. *Cell*, 71(1):107–118, Oct. 1992. doi: 10.1016/0092-8674(92)90270-m.

G. E. Borgstahl, H. E. Parge, M. J. Hickey, M. J. Johnson, M. Boissinot, R. A. Hallewell, J. R. Lepock, D. E. Cabelli, and J. A. Tainer. Human mitochondrial manganese superoxide dismutase polymorphic variant Ile58Thr reduces activity by destabilizing the tetrameric interface. *Biochemistry*, 35(14):4287–4297, Apr. 1996. doi: 10.1021/bi951892w.

M. Born. Zur quantenmechanik der stoßvorgänge. *Eur. Phys. J. A*, 37(12):863–867, Dec. 1926. doi: 10.1007/BF01397477.

M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys.*, 389(20): 457–484, Jan. 1927. doi: 10.1002/andp.19273892002.

S. F. Boys. Electronic wave functions - i. a general method of calculation for the stationary states of any molecular system. *Proc. R. Soc. Lond.*, 200(1063):542–554, Feb. 1950. doi: 10.1098/rspa.1950.0036.

W. H. Bragg and W. L. Bragg. The reflection of x-rays by crystals. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 88(605):428–438, July 1913. doi: 10.1098/rspa.1913.0040.

D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, B. Ni, and S. B. Sinnott. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys. Condens. Matter*, 14(4):783–802, Feb. 2002. doi: 10.1088/0953-8984/14/4/312.

G. Bricogne. [23] bayesian statistical viewpoint on structure determination: Basic concepts and examples. *Methods in Enzymology*, 276:361–423, 1997. doi: 10.1016/S0076-6879(97)76069-5.

P. J. Brown, A. G. Fox, E. N. Maslen, M. A. O'Keefe, and B. T. M. Willis. Intensity of diffracted intensities. In *International Tables for Crystallography*, pages 554–595. International Union of Crystallography, Chester, England, Oct. 2006. doi: 10.1107/97809553602060000600.

A. T. Brünger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, Jan. 1992. doi: 10.1038/355472a0.

A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. *Crystallography & NMR System*: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D Struct. Biol.*, 54(5):905–921, Sept. 1998. doi: 10.1107/s0907444998003254.

A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. M. Krahn, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, G. F. Schroeder, T. Simonson, and G. L. Warren. Crystallography & NMR system, version 1.3. https://cns-online.org/v1.3/, 2010. [Online; accessed 2025-10-05].

G. Bunker, L. Petersson, B. M. Sjöberg, M. Sahlin, M. Chance, B. Chance, and A. Ehrenberg. Extended x-ray absorption fine structure studies on the iron-containing subunit of ribonucleotide reductase from escherichia coli. *Biochemistry*, 26(15):4708–4716, July 1987. doi: 10.1021/bi00389a017.

B. K. Burgess and D. J. Lowe. Mechanism of molybdenum nitrogenase. *Chem. Rev.*, 96 (7):2983–3012, Nov. 1996. doi: 10.1021/cr950055x.

O. Caldararu, F. Manzoni, E. Oksanen, D. T. Logan, and U. Ryde. Refinement of protein structures using a combination of quantum-mechanical calculations with neutron and x-ray crystallographic data. *Acta Crystallogr. D Struct. Biol.*, 75(Pt 4):368–380, Apr. 2019. doi: 10.1107/S205979831900175X.

E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme. A generally applicable atomic-charge dependent london dispersion correction. *J. Chem. Phys.*, 150(15):154122, Apr. 2019. doi: 10.1063/1.5090222.

R. Cammi, B. Mennucci, and J. Tomasi. Fast evaluation of geometries and properties of excited molecules in solution: A tamm-dancoff model with application to 4-dimethylaminobenzonitrile. *J. Phys. Chem. A*, 104(23):5631–5637, June 2000. doi: 10.1021/jp000156l.

L. Cao and U. Ryde. On the difference between additive and subtractive QM/MM calculations. *Front. Chem.*, 6, Apr. 2018. doi: 10.3389/fchem.2018.00089.

L. Cao and U. Ryde. Quantum refinement with multiple conformations: application to the p-cluster in nitrogenase. *Acta Crystallogr. D Struct. Biol.*, 76(Pt 11):1145–1156, Nov. 2020. doi: 10.1107/S2059798320012917.

L. Cao, O. Caldararu, and U. Ryde. Protonation states of homocitrate and nearby residues in nitrogenase studied by computational methods and quantum refinement. *J. Phys. Chem. B*, 121(35):8242–8262, Sept. 2017. doi: 10.1021/acs.jpcb.7b02714.

L. Cao, O. Caldararu, A. C. Rosenzweig, and U. Ryde. Quantum refinement does not support dinuclear copper sites in crystal structures of particulate methane monooxygenase. *Angew. Chem. Int. Ed Engl.*, 57(1):162–166, Jan. 2018. doi: 10.1002/anie.201708977.

L. Cao, O. Caldararu, and U. Ryde. Does the crystal structure of vanadium nitrogenase contain a reaction intermediate? evidence from quantum refinement. *J. Biol. Inorg. Chem.*, 25(6):847–861, Sept. 2020. doi: 10.1007/s00775-020-01813-z.

M. Capdevila-Cortada. Electrifying the Haber–Bosch. *Nat. Catal.*, 2(12):1055–1055, Dec. 2019. doi: 10.1038/s41929-019-0414-4.

S. C. Capelli, H.-B. Bürgi, B. Dittrich, S. Grabowsky, and D. Jayatilaka. Hirshfeld atom refinement. *IUCrJ*, 1(Pt 5):361–379, Sept. 2014. doi: 10.1107/S2052252514014845.

E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.*, 18(5):581–586, Oct. 2008. doi: 10.1016/j.sbi.2008.07.001.

S. I. Chan, V. C.-C. Wang, J. C.-H. Lai, S. S.-F. Yu, P. P.-Y. Chen, K. H.-C. Chen, C.-L. Chen, and M. K. Chan. Redox potentiometry studies of particulate methane monooxygenase: support for a trinuclear copper cluster active site. *Angew. Chem. Int. Ed Engl.*, 46(12):1992–1994, 2007. doi: 10.1002/anie.200604647.

S. I. Chan, W.-H. Chang, S.-H. Huang, H.-H. Lin, and S. S.-F. Yu. Catalytic machinery of methane oxidation in particulate methane monooxygenase (pMMO). *J. Inorg. Biochem.*, 225(111602):111602, Dec. 2021. doi: 10.1016/j.jinorgbio.2021.111602.

S. I. Chan, V. C.-C. Wang, P. P.-Y. Chen, and S. S.-F. Yu. Methane oxidation by the copper methane monooxygenase: Before and after the cryogenic electron microscopy structure of particulate methane monooxygenase from *Methylococcus capsulatus* (bath). *J. Chin. Chem. Soc.*, 69(8):1147–1158, Aug. 2022. doi: 10.1002/jccs.202200166.

W.-H. Chang, H.-H. Lin, I.-K. Tsai, S.-H. Huang, S.-C. Chung, I.-P. Tu, S. S.-F. Yu, and S. I. Chan. Copper centers in the cryo-EM structure of particulate methane monooxygenase reveal the catalytic machinery of methane oxidation. *J. Am. Chem. Soc.*, 143(26): 9922–9932, July 2021. doi: 10.1021/jacs.1c04082.

Y. Cheng, N. Grigorieff, P. A. Penczek, and T. Walz. A primer to single-particle cryo-electron microscopy. *Cell*, 161(3):438–449, Apr. 2015. doi: 10.1016/j.cell.2015.03.050.

S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.*, 9(2):eadf0873, Jan. 2023. doi: 10.1126/sciadv.adf0873.

M. Chruszcz, W. Potrzebowski, M. D. Zimmerman, M. Grabowski, H. Zheng, P. Lasota, and W. Minor. Analysis of solvent content and oligomeric states in protein crystals–does symmetry matter? *Protein Sci.*, 17(4):623–632, Apr. 2008. doi: 10.1110/ps.073360508.

E. Y. D. Chua, J. H. Mendez, M. Rapp, S. L. Ilca, Y. Z. Tan, K. Maruthi, H. Kuang, C. M. Zimanyi, A. Cheng, E. T. Eng, A. J. Noble, C. S. Potter, and B. Carragher. Better, faster, cheaper: Recent advances in cryo-electron microscopy. *Annu. Rev. Biochem.*, 91(1):1–32, June 2022. doi: 10.1146/annurev-biochem-032620-110705.

L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, and K. Morokuma. The ONIOM method and its applications. *Chem. Rev.*, 115(12):5678–5796, June 2015. doi: 10.1021/cr5004419.

D. Cirri, C. Bazzicalupi, U. Ryde, J. Bergmann, F. Binacchi, A. Nocentini, A. Pratesi, P. Gratteri, and L. Messori. Computationally enhanced x-ray diffraction analysis of a gold(III) complex interacting with the human telomeric DNA g-quadruplex. unravelling non-unique ligand positioning. *Int. J. Biol. Macromol.*, 211:506–513, June 2022. doi: 10.1016/j.ijbiomac.2022.05.033.

C. M. Clemente, L. Capece, and M. A. Martí. Best practices on QM/MM simulations of biological systems. *J. Chem. Inf. Model.*, 63(9):2609–2627, May 2023. doi: 10.1021/acs.jcim.2c01522.

A. H. Compton. The distribution of the electrons in atoms. *Nature*, 95(2378):343–344, May 1915. doi: 10.1038/095343b0.

J. A. Cotruvo, Jr, T. A. Stich, R. D. Britt, and J. Stubbe. Mechanism of assembly of the dimanganese-tyrosyl radical cofactor of class ib ribonucleotide reductase: enzymatic generation of superoxide is required for tyrosine oxidation via a Mn(III)Mn(IV) intermediate. *J. Am. Chem. Soc.*, 135(10):4027–4039, Mar. 2013. doi: 10.1021/ja312457t.

K. Cowtan. *Phase problem in X-ray crystallography, and its solution*. John Wiley & Sons, Ltd, Chichester, May 2003. doi: 10.1038/npg.els.0002722.

D. T. Cromer and J. B. Mann. X-ray scattering factors computed from numerical Hartree–Fock wave functions. *Acta Crystallogr. A*, 24(2):321–324, Mar. 1968. doi: 10.1107/S0567739468000550.

CryoSPARC. Contrast in cryo-em. https://guide.cryosparc.com/cryo-em-foundations/image-formation/contrast-in-cryo-em, 2025. [Online; accessed 2025-09-30].

M. A. Culpepper and A. C. Rosenzweig. Architecture and active site of particulate methane monooxygenase. *Crit. Rev. Biochem. Mol. Biol.*, 47(6):483–492, Nov. 2012. doi: 10.3109/10409238.2012.697865.

L. de Broglie. On the theory of quanta. https://fondationlouisdebroglie.org/LDB-oeuvres/De_Broglie_Kracklauer.pdf, 1924. [Online; accessed 2025-09-30].

L. de Broglie. Recherches sur la théorie des quanta. *Ann. Phys. (Paris)*, 10(3):22–128, 1925. doi: 10.1051/anphys/192510030022.

P. Debye. Interferenz von röntgenstrahlen und wärmebewegung. *Ann. Phys.*, 348(1):49–92, Jan. 1913. doi: 10.1002/andp.19133480105.

S. K. Dhar and D. K. St Clair. Manganese superoxide dismutase regulation and cancer. *Free Radic. Biol. Med.*, 52(11-12):2209–2222, Apr. 2012. doi: 10.1016/j.freeradbiomed.2012.03.009.

P. A. M. Dirac. On the theory of quantum mechanics. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 112(762):661–677, Oct. 1926. doi: 10.1098/rspa.1926.0133.

P. A. M. Dirac. A new notation for quantum mechanics. *Math. Proc. Camb. Philos. Soc.*, 35(3):416–418, July 1939. doi: 10.1017/S0305004100021162.

O. Einsle. Nitrogenase FeMo cofactor: an atomic structure in three simple steps. *J. Biol. Inorg. Chem.*, 19(6):737–745, Aug. 2014. doi: 10.1007/s00775-014-1116-7.

H. Eklund, U. Uhlin, M. Färnegårdh, D. T. Logan, and P. Nordlund. Structure and function of the radical enzyme ribonucleotide reductase. *Prog. Biophys. Mol. Biol.*, 77(3):177–268, Nov. 2001. doi: 10.1016/s0079-6107(01)00014-1.

N. Elango, R. Radhakrishnan, W. A. Froland, B. J. Wallar, C. A. Earhart, J. D. Lipscomb, and D. H. Ohlendorf. Crystal structure of the hydroxylase component of methane monooxygenase from methylosinus trichosporium OB3b. *Protein Sci.*, 6(3):556–568, Mar. 1997. doi: 10.1002/pro.5560060305.

R. Eliasson, M. Fontecave, H. Jörnvall, M. Krook, E. Pontis, and P. Reichard. The anaerobic ribonucleoside triphosphate reductase from escherichia coli requires s-adenosylmethionine as a cofactor. *Proc. Natl. Acad. Sci. U. S. A.*, 87(9):3314–3318, May 1990. doi: 10.1073/pnas.87.9.3314.

R. A. Engh and R. Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallogr. A*, 47(4):392–400, July 1991. doi: 10.1107/S0108767391001071.

R. A. Engh and R. Huber. Structure quality and target parameters. In *International Tables for Crystallography*, pages 474–484. International Union of Crystallography, Chester, England, Apr. 2012. doi: 10.1107/97809553602060000857.

J. W. Erisman, M. A. Sutton, J. Galloway, Z. Klimont, and W. Winiwarter. How a century of ammonia synthesis changed the world. *Nat. Geosci.*, 1(10):636–639, Oct. 2008. doi: 10.1038/ngeo325.

P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Ann. Phys.*, 369(3):253–287, Jan. 1921. doi: 10.1002/andp.19213690304.

B. Falkner and G. F. Schröder. Cross-validation in cryo-EM-based structural modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 110(22):8930–8935, May 2013. doi: 10.1073/pnas.1119041110.

T. D. Fenn and M. J. Schnieders. Polarizable atomic multipole x-ray refinement: weighting schemes for macromolecular diffraction. *Acta Crystallogr. D Struct. Biol.*, 67(Pt 11):957–965, Nov. 2011. doi: 10.1107/S0907444911039060.

T. D. Fenn, M. J. Schnieders, and A. T. Brunger. A smooth and differentiable bulk-solvent model for macromolecular diffraction. *Acta Crystallogr. D Struct. Biol.*, 66(Pt 9):1024–1031, Sept. 2010. doi: 10.1107/S0907444910031045.

M. J. Field, P. A. Bash, and M. Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.*, 11(6):700–733, July 1990. doi: 10.1002/jcc.540110605.

V. Fock. Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems. *Eur. Phys. J. A*, 61(1-2):126–148, Jan. 1930. doi: 10.1007/BF01340294.

J. Frank. *Three-dimensional electron microscopy of macromolecular assemblies*. Oxford University Press, New York, NY, Sept. 2005. ISBN 9780199893416. doi: 10.1093/acprof:oso/9780195182187.001.0001.

T. Fransson, M. G. Delcey, I. E. Brumboiu, M. Hodecker, X. Li, Z. Rinkevicius, A. Dreuw, Y. M. Rhee, and P. Norman. *Computational Chemistry from Laptop to HPC : A notebook exploration of quantum chemistry*. KTH Royal Institute of Technology, Stockholm, Aug. 2022. doi: 10.30746/978-91-988114-0-7.

F. Furche, R. Ahlrichs, C. Hättig, W. Klopper, M. Sierka, and F. Weigend. Turbomole. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 4(2):91–100, Mar. 2014. doi: 10.1002/wcms.1162.

D. García-Aldea and J. E. Alvarellos. Kinetic energy density study of some representative semilocal kinetic energy functionals. *J. Chem. Phys.*, 127(14):144109, Oct. 2007. doi: 10.1063/1.2774974.

J. A. Gaunt. A theory of hartree's atomic fields. *Math. Proc. Camb. Philos. Soc.*, 24(2): 328–342, Apr. 1928. doi: 10.1017/S0305004100015851.

S. Giannini, P. M. Martinez, A. Semmeq, J. P. Galvez, A. Piras, A. Landi, D. Padula, J. G. Vilhena, J. Cerezo, and G. Prampolini. JOYCE3.0: A general protocol for the specific parametrization of accurate intramolecular quantum mechanically derived force fields. *J. Chem. Theory Comput.*, 21(6):3156–3175, Mar. 2025. doi: 10.1021/acs.jctc.5c00010.

S. Grabowsky, A. Genoni, S. P. Thomas, and D. Jayatilaka. The advent of quantum crystallography: Form and structure factors from quantum mechanics for advanced structure refinement and wavefunction fitting. In *Structure and Bonding*, Structure and bonding, pages 65–144. Springer International Publishing, Cham, 2020. ISBN 978-3-030-64746-9. doi: 10.1007/430_2020_62.

D. W. Green, V. M. Ingram, and M. F. Perutz. The structure of haemoglobin - IV. sign determination by the isomorphous replacement method. *Proc. R. Soc. Lond.*, 225(1162): 287–307, Sept. 1954. doi: 10.1098/rspa.1954.0203.

B. L. Greene, G. Kang, C. Cui, M. Bennati, D. G. Nocera, C. L. Drennan, and J. Stubbe. Ribonucleotide reductases: Structure, chemistry, and metabolism suggest new therapeutic targets. *Annu. Rev. Biochem.*, 89(1):45–75, June 2020. doi: 10.1146/annurev-biochem-013118-111843.

R. W. Grosse-Kunstleve and P. D. Adams. On the handling of atomic anisotropic displacement parameters. *J. Appl. Crystallogr.*, 35(4):477–480, Aug. 2002. doi: 10.1107/S0021889802008580.

R. W. Grosse-Kunstleve, N. K. Sauter, N. W. Moriarty, and P. D. Adams. The*Computational* Crystallography Toolbox: crystallographic algorithms in a reusable software framework. *J. Appl. Crystallogr.*, 35(1):126–136, Feb. 2002. doi: 10.1107/S0021889801017824.

R. W. Grosse-Kunstleve, B. Wong, M. Mustyakimov, and P. D. Adams. Exact direct-space asymmetric units for the 230 crystallographic space groups. *Acta Crystallogr. A*, 67(Pt 3): 269–275, May 2011. doi: 10.1107/S0108767311007008.

G. G. Hall. The molecular orbital theory of chemical valency VIII. a method of calculating ionization potentials. *Proc. R. Soc. Lond.*, 205(1083):541–552, Mar. 1951. doi: 10.1098/rspa.1951.0048.

N. K. Hansen and P. Coppens. Testing aspherical atom refinements on small-molecule data sets. *Acta Crystallogr. A*, 34(6):909–921, Nov. 1978. doi: 10.1107/S0567739478001886.

R. S. Hanson and T. E. Hanson. Methanotrophic bacteria. *Microbiol. Rev.*, 60(2):439–471, June 1996. doi: 10.1128/mr.60.2.439-471.1996.

D. F. Harris, D. A. Lukoyanov, H. Kallas, C. Trncik, Z.-Y. Yang, P. Compton, N. Kelleher, O. Einsle, D. R. Dean, B. M. Hoffman, and L. C. Seefeldt. Mo-, v-, and fe-nitrogenases use a universal eight-electron reductive-elimination mechanism to achieve N2 reduction. *Biochemistry*, 58(30):3293–3301, July 2019. doi: 10.1021/acs.biochem.9b00468.

P. C. Hart, M. Mao, A. L. P. de Abreu, K. Ansenberger-Fricano, D. N. Ekoue, D. Ganini, A. Kajdacsy-Balla, A. M. Diamond, R. D. Minshall, M. E. L. Consolaro, J. H. Santos, and M. G. Bonini. MnSOD upregulation sustains the warburg effect via mitochondrial ROS and AMPK-dependent signalling in cancer. *Nat. Commun.*, 6(1):6053, Feb. 2015. doi: doi.org/10.1038/ncomms7053.

D. R. Hartree. The wave mechanics of an atom with a non-coulomb central field. part II. some results and discussion. *Math. Proc. Camb. Philos. Soc.*, 24(1):111–132, Jan. 1928a. doi: 10.1017/S0305004100011920.

D. R. Hartree. The wave mechanics of an atom with a non-coulomb central field. part II. some results and discussion. *Math. Proc. Camb. Philos. Soc.*, 24(1):111–132, Jan. 1928b. doi: 10.1017/S0305004100011920.

D. R. Hartree and W. Hartree. Self-consistent field, with exchange, for beryllium. *Proc. R. Soc. Lond.*, 150(869):9–33, May 1935. doi: 10.1098/rspa.1935.0085.

W. Heisenberg. Mehrkörperproblem und resonanz in der quantenmechanik. *Eur. Phys. J. A*, 38(6-7):411–426, June 1926. doi: 10.1007/BF01397160.

W. A. Hendrickson. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, 254(5028):51–58, Oct. 1991. doi: 10.1126/science.1925561.

W. A. Hendrickson, J. R. Horton, and D. M. LeMaster. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (mad): a vehicle for direct determination of three-dimensional structure. *EMBO J.*, 9(5):1665–1672, May 1990. doi: 10.1002/j.1460-2075.1990.tb08287.x.

H.-P. Hersleth and K. K. Andersson. How different oxidation states of crystalline myoglobin are influenced by x-rays. *Biochim. Biophys. Acta*, 1814(6):785–796, June 2011. doi: 10.1016/j.bbapap.2010.07.019.

F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoret. Chim. Acta*, 44(2):129–138, 1977. doi: 10.1007/BF00549096.

A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter*, 29 (27):273002, July 2017. doi: 10.1088/1361-648X/aa680e.

A. Hofer, M. Crona, D. T. Logan, and B.-M. Sjöberg. DNA building blocks: keeping control of manufacture. *Crit. Rev. Biochem. Mol. Biol.*, 47(1):50–63, Jan. 2012. doi: 10.3109/10409238.2011.630372.

G. Hofer, L. Wang, L. Pacoste, P. Hager, A. Fonjallaz, L. Williams, E. Scaletti Hutchinson, M. Di Palma, P. Stenmark, J. Worral, R. A. Steiner, H. Xu, and X. Zou. Continuous serial electron diffraction for high quality protein structures. *bioRxivorg*, Sept. 2025. doi: 10.1101/2025.09.14.676192.

B. M. Hoffman, D. Lukoyanov, Z.-Y. Yang, D. R. Dean, and L. C. Seefeldt. Mechanism of nitrogen fixation by nitrogenase: the next stage. *Chem. Rev.*, 114(8):4041–4062, Apr. 2014. doi: 10.1021/cr400641x.

M. Högbom, M. Galander, M. Andersson, M. Kolberg, W. Hofbauer, G. Lassmann, P. Nordlund, and F. Lendzian. Displacement of the tyrosyl radical cofactor in ribonucleotide reductase obtained by single-crystal high-field EPR and 1.4-a x-ray data. *Proc. Natl. Acad. Sci. U. S. A.*, 100(6):3209–3214, Mar. 2003. doi: 10.1073/pnas.0536684100.

P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, Nov. 1964. doi: 10.1103/PhysRev.136.B864.

A. Hoser and A. Ø. Madsen. Models of thermal motion in small-molecule crystallography. *IUCrJ*, 12(Pt 4):421–434, July 2025. doi: 10.1107/S2052252525004361.

Y.-W. Hsiao, T. Drakenberg, and U. Ryde. NMR structure determination of proteins supplemented by quantum chemical calculations: detailed structure of the ca2+ sites in the EGF34 fragment of protein S. *J. Biomol. NMR*, 31(2):97–114, Feb. 2005. doi: 10.1007/s10858-004-6729-7.

Y.-W. Hsiao, Y. Tao, J. E. Shokes, R. A. Scott, and U. Ryde. EXAFS structure refinement supplemented by computational chemistry. *Phys. Rev. B Condens. Matter Mater. Phys.*, 74(21), Dec. 2006. doi: 10.1103/PhysRevB.74.214101.

Y.-W. Hsiao, E. Sanchez-Garcia, M. Doerr, and W. Thiel. Quantum refinement of protein structures: implementation and application to the red fluorescent protein DsRed.M1. *J. Phys. Chem. B*, 114(46):15413–15423, Nov. 2010. doi: 10.1021/jp108095n.

L. Hu and U. Ryde. Comparison of methods to obtain force-field parameters for metal sites. *J. Chem. Theory Comput.*, 7(8):2452–2463, Aug. 2011. doi: 10.1021/ct100725a.

IUCr. Definition of a crystal. *Acta Crystallogr. A Found. Adv.*, 48(A):922–946, 1992.

IUCr. Statistical descriptors in crystallography. https://www.iucr.org/resources/commissions/crystallographic-nomenclature/statdes/refine.html, Sept. 1996. [Online; accessed 2025-10-01].

IUCr. Definition of a crystal. https://dictionary.iucr.org/Crystal, 2021. [Online; accessed 2025-09-02].

A. Jack and M. Levitt. Refinement of large structures by simultaneous minimization of energy and *R* factor. *Acta Crystallogr. A*, 34(6):931–935, Nov. 1978. doi: 10.1107/S0567739478001904.

F. Jensen. *Introduction to computational chemistry*. John Wiley & Sons, Nashville, TN, 3 edition, Feb. 2017. ISBN 9781118825990.

H. Jiang, Y. Chen, P. Jiang, C. Zhang, T. J. Smith, J. C. Murrell, and X.-H. Xing. Methanotrophs: Multifunctional bacteria with promising applications in environmental bioengineering. *Biochem. Eng. J.*, 49(3):277–288, May 2010. doi: 10.1016/j.bej.2010.01.003.

J. S. Jiang and A. T. Brünger. Protein hydration observed by x-ray diffraction. solvation properties of penicillopepsin and neuraminidase crystal structures. *J. Mol. Biol.*, 243(1):100–115, Oct. 1994. doi: 10.1006/jmbi.1994.1633.

S. G. Johnson. The NLopt nonlinear-optimization package. `https://github.com/stevengj/nlopt`, 2007.

T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A*, 47(2):110–119, Mar. 1991. doi: 10.1107/S0108767390010224.

A. P. Joseph, I. Lagerstedt, A. Patwardhan, M. Topf, and M. Winn. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.*, 199(1):12–26, July 2017. doi: 10.1016/j.jsb.2017.05.007.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. doi: 10.1038/s41586-021-03819-2.

H. Kamberaj. Basis set functions. In *Scientific Computation*, Scientific computation, pages 31–54. Springer Nature Switzerland, Cham, 2023. doi: 10.1007/978-3-031-34839-6_2.

J. Kim and D. C. Rees. Structural models for the metal centers in the nitrogenase molybdenum-iron protein. *Science*, 257(5077):1677–1682, Sept. 1992. doi: 10.1126/science.1529354.

C. Kittel. *Introduction to solid state physics*. John Wiley and Sons (WIE), Brisbane, QLD, Australia, 8 edition, Dec. 2004. ISBN 9780471415268.

F. Kleemiss, O. V. Dolomanov, M. Bodensteiner, N. Peyerimhoff, L. Midgley, L. J. Bourhis, A. Genoni, L. A. Malaspina, D. Jayatilaka, J. L. Spencer, F. White, B. Grundkötter-Stock, S. Steinhauer, D. Lentz, H. Puschmann, and S. Grabowsky. Accurate crystal structures and chemical properties from NoSpherA2. *Chem. Sci.*, 12(5):1675–1692, Nov. 2020. doi: 10.1039/D0SC05526C.

G. J. Kleywegt and T. A. Jones. Databases in protein crystallography. *Acta Crystallogr. D Struct. Biol.*, 54(6):1119–1131, Nov. 1998. doi: 10.1107/S0907444998007100.

J. Kobus. A finite difference Hartree–Fock program for atoms and diatomic molecules. *Comput. Phys. Commun.*, 184(3):799–811, Mar. 2013. doi: 10.1016/j.cpc.2012.09.033.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov. 1965. doi: 10.1103/PhysRev.140.A1133.

W. Kohn, A. D. Becke, and R. G. Parr. Density functional theory of electronic structure. *J. Phys. Chem.*, 100(31):12974–12980, Jan. 1996. doi: 10.1021/jp960669l.

C. W. Koo, F. J. Tucci, Y. He, and A. C. Rosenzweig. Recovery of particulate methane monooxygenase structure and activity in a lipid bilayer. *Science*, 375(6586):1287–1291, Mar. 2022. doi: 10.1126/science.abm3282.

W. Kühlbrandt. Biochemistry. the resolution revolution. *Science*, 343(6178):1443–1444, Mar. 2014. doi: 10.1126/science.1251652.

M. Laue. Eine quantitative prüfung der theorie für die interferenzerscheinungen bei röntgenstrahlen. *Ann. Phys.*, 346(10):989–1002, Jan. 1913. doi: 10.1002/andp.19133461005.

M. Laue. Concerning the Detection of X-ray Interferences. https://www.nobelprize.org/prizes/physics/1914/laue/lecture/, 1920. [Online; accessed 2025-09-06].

C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B Condens. Matter*, 37(2): 785–789, Jan. 1988. doi: 10.1103/PhysRevB.37.785.

I. Leven, H. Hao, S. Tan, X. Guan, K. A. Penrod, D. Akbarian, B. Evangelisti, M. J. Hossain, M. M. Islam, J. P. Koski, S. Moore, H. M. Aktulga, A. C. T. van Duin, and T. Head-Gordon. Recent advances for improving the accuracy, transferability, and efficiency of reactive force fields. *J. Chem. Theory Comput.*, 17(6):3237–3251, June 2021. doi: 10.1021/acs.jctc.1c00118.

M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, 46(2):269–279, Dec. 1969. doi: 10.1016/0022-2836(69)90421-5.

R. L. Lieberman and A. C. Rosenzweig. The quest for the particulate methane monooxygenase active site. *Dalton Trans.*, (21):3390–3396, Nov. 2005a. doi: 10.1039/B506651D.

R. L. Lieberman and A. C. Rosenzweig. Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane. *Nature*, 434(7030):177–182, Mar. 2005b. doi: 10.1038/nature03311.

R. L. Lieberman, D. B. Shrestha, P. E. Doan, B. M. Hoffman, T. L. Stemmler, and A. C. Rosenzweig. 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.*, 100(7):3820–3825, Apr. 2003. doi: 10.1073/pnas.0536703100.

D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, and P. D. Adams. Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallogr. D Struct. Biol.*, 75(Pt 10):861–877, Oct. 2019. doi: 10.1107/S2059798319011471.

D. Liebschner, N. W. Moriarty, B. K. Poon, and P. D. Adams. In situ ligand restraints from quantum-mechanical methods. *Acta Crystallogr. D Struct. Biol.*, 79(Pt 2):100–110, Feb. 2023. doi: 10.1107/S2059798323000025.

A. Liljas, L. Liljas, G. Lindblom, P. Nissen, M. Kjeldgaard, and M. Ash. *Textbook Of Structural Biology (Second Edition)*. Series In Structural Biology. World Scientific Publishing Company, 2016. ISBN 9789813142497. doi: 10.1142/6620. URL `https://www.worldscientific.com/worldscibooks/10.1142/6620`.

H. Lin and D. G. Truhlar. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.*, 117(2):185–199, Feb. 2007. doi: 10.1007/s00214-006-0143-z.

I. Lobato and D. Van Dyck. An accurate parameterization for scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints. *Acta Crystallogr. A Found. Adv.*, 70(6):636–649, Nov. 2014. doi: 10.1107/S205327331401643X.

D. T. Logan, X. D. Su, A. Aberg, K. Regnström, J. Hajdu, H. Eklund, and P. Nordlund. Crystal structure of reduced protein R2 of ribonucleotide reductase: the structural basis for oxygen activation at a dinuclear iron site. *Structure*, 4(9):1053–1064, Sept. 1996. doi: 10.1016/S0969-2126(96)00112-8.

P. J. Loll. Membrane proteins, detergents and crystals: what is the state of the art? *Acta Crystallogr. F Struct. Biol. Commun.*, 70(Pt 12):1576–1583, Dec. 2014. doi: 10.1107/S2053230X14025035.

N. London, D. K. Limbu, M. O. Faruque, F. A. Shakib, and M. R. Momeni. DL_POLY quantum 2.1: A suite of real-time path integral methods for the simulation of dynamical properties and vibrational spectra. *J. Phys. Chem. A*, 129(18):4015–4028, May 2025. doi: 10.1021/acs.jpca.4c08644.

F. Long, R. A. Nicholls, P. Emsley, S. Gra□ulis, A. Merkys, A. Vaitkus, and G. N. Murshudov. AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr. D Struct. Biol.*, 73(Pt 2):112–122, Feb. 2017. doi: 10.1107/S2059798317000067.

T. Lovell, J. Li, T. Liu, D. A. Case, and L. Noodleman. FeMo cofactor of nitrogenase: A density functional study of states MN, MOX, MR, and MI. *J. Am. Chem. Soc.*, 123(49): 12392–12410, Dec. 2001. doi: 10.1021/ja011860y.

Y. Lu, M. R. Farrow, P. Fayon, A. J. Logsdail, A. A. Sokol, C. R. A. Catlow, P. Sherwood, and T. W. Keal. Open-source, python-based redevelopment of the ChemShell multiscale QM/MM environment. *J. Chem. Theory Comput.*, 15(2):1317–1328, Feb. 2019. doi: 10.1021/acs.jctc.8b01036.

D. A. Lukoyanov, D. F. Harris, Z.-Y. Yang, A. Pérez-González, D. R. Dean, L. C. Seefeldt, and B. M. Hoffman. The one-electron reduced active-site FeFe-cofactor of fe-nitrogenase contains a hydride bound to a formally oxidized metal-ion core. *Inorg. Chem.*, 61(14): 5459–5464, Apr. 2022. doi: 10.1021/acs.inorgchem.2c00180.

K. J. M. Lundgren, L. Cao, M. Torbjörnsson, E. D. Hedegård, and U. Ryde. The CuB site in particulate methane monooxygenase may be used to produce hydrogen peroxide. *Dalton Trans.*, 54(8):3141–3156, Feb. 2025. doi: 10.1039/d4dt03301a.

D. Lundin, S. Gribaldo, E. Torrents, B.-M. Sjöberg, and A. M. Poole. Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. *BMC Evol. Biol.*, 10(1):383, Dec. 2010. doi: 10.1186/1471-2148-10-383.

V. Y. Lunin, P. V. Afonine, and A. G. Urzhumtsev. Likelihood-based refinement. i. irremovable model errors. *Acta Crystallogr. A*, 58(3):270–282, May 2002. doi: 10.1107/S0108767302001046.

V. Luzzati. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallogr.*, 5(6):802–810, Nov. 1952. doi: 10.1107/S0365110X52002161.

A. D. Mackerell, Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.*, 25(13):1584–1604, Oct. 2004. doi: 10.1002/jcc.20082.

J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. Ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, Aug. 2015. doi: 10.1021/acs.jctc.5b00255.

M. Martinho, D. W. Choi, A. A. Dispirito, W. E. Antholine, J. D. Semrau, and E. Münck. Mössbauer studies of the membrane-associated methane monooxygenase from methylococcus capsulatus bath: evidence for a diiron center. *J. Am. Chem. Soc.*, 129(51):15783–15785, Dec. 2007. doi: 10.1021/ja077682b.

B. W. Matthews. Solvent content of protein crystals. *J. Mol. Biol.*, 33(2):491–497, Apr. 1968. doi: 10.1016/0022-2836(68)90205-2.

A. J. McCoy. Liking likelihood. *Acta Crystallogr. D Struct. Biol.*, 60(Pt 12 Pt 1):2169–2183, Dec. 2004. doi: 10.1107/S0907444904016038.

A. J. McCoy, M. D. Sammito, and R. J. Read. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallogr. D Struct. Biol.*, 78(Pt 1):1–13, Jan. 2022. doi: 10.1107/S2059798321012122.

M. Merkx, D. A. Kopp, M. H. Sazinsky, J. L. Blazyk, J. Müller, and S. J. Lippard. Dioxygen activation and methane hydroxylation by soluble methane monooxygenase: A tale of two irons and three proteins. *Angew. Chem. Int. Ed Engl.*, 40(15):2782–2807, Aug. 2001. doi: 10.1002/1521-3773(20010803)40:15<2782::AID-ANIE2782>3.0.CO;2-P.

K. M. Merz, Jr. Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.*, 47(9):2804–2811, Sept. 2014. doi: 10.1021/ar5001023.

O. Mikhailovskii, S. A. Izmailov, Y. Xue, D. A. Case, and N. R. Skrynnikov. X-ray crystallography module in MD simulation program amber 2023. refining the models of protein crystals. *J. Chem. Inf. Model.*, 64(1):18–25, Jan. 2024. doi: 10.1021/acs.jcim.3c01531.

J. L. S. Milne, M. J. Borgnia, A. Bartesaghi, E. E. H. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, and S. Subramaniam. Cryo-electron microscopy–a primer for the non-microscopist. *FEBS J.*, 280(1):28–45, Jan. 2013. doi: 10.1111/febs.12078.

P. C. Moews and R. H. Kretsinger. Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference fourier analysis. *J. Mol. Biol.*, 91 (2):201–225, Jan. 1975. doi: 10.1016/0022-2836(75)90160-6.

N. W. Moriarty, R. W. Grosse-Kunstleve, and P. D. Adams. electronic ligand builder and optimization workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D Struct. Biol.*, 65(Pt 10):1074–1080, Oct. 2009. doi: 10.1107/S0907444909029436.

N. W. Moriarty, D. E. Tronrud, P. D. Adams, and P. A. Karplus. A new default restraint library for the protein backbone in phenix: a conformation-dependent geometry goes mainstream. *Acta Crystallogr. D Struct. Biol.*, 72(Pt 1):176–179, Jan. 2016. doi: 10.1107/S2059798315022408.

N. W. Moriarty, P. A. Janowski, J. M. Swails, H. Nguyen, J. S. Richardson, D. A. Case, and P. D. Adams. Improved chemistry restraints for crystallographic refinement by integrating the amber force field into phenix. *Acta Crystallogr. D Struct. Biol.*, 76(Pt 1): 51–62, Jan. 2020. doi: 10.1107/S2059798319015134.

N. W. Moriarty, D. A. Case, D. Liebschner, and P. D. Adams. Validated ligand geometries for macromolecular refinement restraints and molecular mechanics force fields. *bioRxivorg*, Aug. 2025. doi: 10.1101/2025.08.01.668229.

N. F. Mott. The scattering of electrons by atoms. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 127(806):658–665, June 1930. doi: 10.1098/rspa.1930.0082.

T. Mueller, A. Hernandez, and C. Wang. Machine learning for interatomic potential models. *J. Chem. Phys.*, 152(5):050902, Feb. 2020. doi: 10.1063/1.5126336.

R. B. Murphy, D. M. Philipp, and R. A. Friesner. Frozen orbital QM/MM methods for density functional theory. *Chem. Phys. Lett.*, 321(1-2):113–120, Apr. 2000. doi: 10.1016/S0009-2614(00)00289-X.

G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Struct. Biol.*, 53(Pt 3): 240–255, May 1997. doi: 10.1107/S0907444996012255.

G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Struct. Biol.*, 67(Pt 4):355–367, Apr. 2011. doi: 10.1107/S0907444911001314.

F. Neese. A critical evaluation of DFT, including time-dependent DFT, applied to bioinorganic chemistry. *J. Biol. Inorg. Chem.*, 11(6):702–711, Sept. 2006. doi: 10.1007/s00775-006-0138-1.

F. Neese. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(1):73–78, Jan. 2012. doi: 10.1002/wcms.81.

F. Neese. Software update: The ORCA program system—version 6.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 15(2), Mar. 2025. doi: 10.1002/wcms.70019.

H. H. Nguyen, A. K. Shiemke, S. J. Jacobs, B. J. Hales, M. E. Lidstrom, and S. I. Chan. The nature of the copper ions in the membranes containing the particulate methane monooxygenase from methylococcus capsulatus (bath). *J. Biol. Chem.*, 269(21):14995–15005, May 1994. doi: 10.1016/S0021-9258(17)36565-1.

K. Nilsson and U. Ryde. Protonation status of metal-bound ligands can be determined by quantum refinement. *J. Inorg. Biochem.*, 98(9):1539–1546, Sept. 2004. doi: 10.1016/j.jinorgbio.2004.06.006.

K. Nilsson, D. Lecerof, E. Sigfridsson, and U. Ryde. An automatic method to generate force-field parameters for hetero-compounds. *Acta Crystallogr. D Struct. Biol.*, 59(2):274–289, Feb. 2003. doi: 10.1107/s0907444902021431.

L. Noodleman. Valence bond description of antiferromagnetic coupling in transition metal dimers. *J. Chem. Phys.*, 74(10):5737–5743, May 1981. doi: 10.1063/1.440939.

P. Nordlund and H. Eklund. Structure and function of the escherichia coli ribonucleotide reductase protein R2. *J. Mol. Biol.*, 232(1):123–164, July 1993. doi: 10.1006/jmbi.1993.1374.

P. Nordlund and P. Reichard. Ribonucleotide reductases. *Annu. Rev. Biochem.*, 75(1):681–706, 2006. doi: 10.1146/annurev.biochem.75.103004.142443.

P. Nordlund, B. M. Sjöberg, and H. Eklund. Three-dimensional structure of the free radical protein of ribonucleotide reductase. *Nature*, 345(6276):593–598, June 1990. doi: 10.1038/345593a0.

P. Novelli, G. Meanti, P. J. Buigues, L. Rosasco, M. Parrinello, M. Pontil, and L. Bonati. Fast and fourier features for transfer learning of interatomic potentials. *Npj Comput. Mater.*, 11(1):293, Sept. 2025. doi: 10.1038/s41524-025-01779-z.

L. Pacoste. *Developing electron diffraction methods to probe oxidation states in metalloenzymes*. Stockholm University, 2025. ISBN 978-91-8107-208-2.

N. S. Pannu, G. N. Murshudov, E. J. Dodson, and R. J. Read. Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr. D Struct. Biol.*, 54(6):1285–1294, Nov. 1998. doi: 10.1107/s0907444998004119.

A. L. Patterson. A fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.*, 46(5):372–376, Sept. 1934. doi: 10.1103/PhysRev.46.372.

L.-M. Peng. Electron scattering factors of ions and their parameterization. *Acta Crystallogr. A*, 54(4):481–485, July 1998. doi: 10.1107/S0108767398001901.

W. Peng, Z. Wang, Q. Zhang, S. Yan, and B. Wang. Unraveling the valence state and reactivity of copper centers in membrane-bound particulate methane monooxygenase. *J. Am. Chem. Soc.*, 145(46):25304–25317, Nov. 2023. doi: 10.1021/jacs.3c08834.

J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, Oct. 1996. doi: 10.1103/PhysRevLett.77.3865.

J. P. Perdew, A. Ruzsinszky, L. A. Constantin, J. Sun, and G. I. Csonka. Some fundamental issues in ground-state density functional theory: A guide for the perplexed. *J. Chem. Theory Comput.*, 5(4):902–908, Apr. 2009. doi: 10.1021/ct800531s.

S. J. Pitman, A. K. Evans, R. T. Ireland, F. Lempriere, and L. K. McKemmish. Benchmarking basis sets for density functional theory thermochemistry calculations: Why unpolarized basis sets and the polarized 6-311G family should be avoided. *J. Phys. Chem. A*, 127(48):10295–10306, Nov. 2023. doi: 10.1021/acs.jpca.3c05573.

Proteopedia. Free R. https://proteopedia.org/wiki/index.php/Free_R, 2025. [Online; accessed 2025-09-16].

A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods*, 14(3):290–296, Mar. 2017. doi: 10.1038/nmeth.4169.

A. Punjani, H. Zhang, and D. J. Fleet. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods*, 17(12):1214–1221, Dec. 2020. doi: 10.1038/s41592-020-00990-8.

R. J. Read. Structure-factor probabilities for related structures. *Acta Crystallogr. A*, 46(11):900–912, Nov. 1990. doi: 10.1107/S0108767390005529.

R. J. Read. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D Struct. Biol.*, 57(10):1373–1382, Oct. 2001. doi: 10.1107/S0907444901012471.

R. J. Read, P. D. Adams, W. B. Arendall, 3rd, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart. A new generation of

crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, Oct. 2011. doi: 10.1016/j.str.2011.08.006.

L. Reimer and H. Kohl. *Transmission electron microscopy*. SPringer series in optical sciences. Springer, New York, NY, 5 edition, Aug. 2008. ISBN 978-1-4419-2308-0. doi: 10.1007/978-0-387-40093-8.

G. Rhodes. *Crystallography made crystal clear*. Complementary Science. Academic Press, San Diego, CA, 3 edition, Feb. 2006. ISBN 9780125870733.

S. Y. Ro, L. F. Schachner, C. W. Koo, R. Purohit, J. P. Remis, G. E. Kenney, B. W. Liauw, P. M. Thomas, S. M. Patrie, N. L. Kelleher, and A. C. Rosenzweig. Native top-down mass spectrometry provides insights into the copper centers of membrane-bound methane monooxygenase. *Nat. Commun.*, 10(1):2675, June 2019. doi: 10.1038/s41467-019-10590-6.

C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.*, 23(2):69–89, Apr. 1951. doi: 10.1103/RevModPhys.23.69.

A. C. Rosenzweig. The metal centres of particulate methane mono-oxygenase. *Biochem. Soc. Trans.*, 36(Pt 6):1134–1137, Dec. 2008. doi: 10.1042/BST0361134.

A. C. Rosenzweig, C. A. Frederick, S. J. Lippard, and P. Nordlund. Crystal structure of a bacterial non-haem iron hydroxylase that catalyses the biological oxidation of methane. *Nature*, 366(6455):537–543, Dec. 1993. doi: 10.1038/366537a0.

M. O. Ross and A. C. Rosenzweig. A tale of two methane monooxygenases. *J. Biol. Inorg. Chem.*, 22(2-3):307–319, Apr. 2017. doi: 10.1007/s00775-016-1419-y.

M. O. Ross, F. MacMillan, J. Wang, A. Nisthal, T. J. Lawton, B. D. Olafson, S. L. Mayo, A. C. Rosenzweig, and B. M. Hoffman. Particulate methane monooxygenase contains only mononuclear copper centers. *Science*, 364(6440):566–570, May 2019. doi: 10.1126/science.aav2572.

M. G. Rossmann. The molecular replacement method. *Acta Crystallogr. A*, 46(2):73–82, Feb. 1990. doi: 10.1107/S0108767389009815.

M. G. Rossmann and D. M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.*, 15(1):24–31, Jan. 1962. doi: 10.1107/S0365110X62000067.

M. G. Rossmann, R. Bernal, and S. V. Pletnev. Combining electron microscopic with x-ray crystallographic structures. *J. Struct. Biol.*, 136(3):190–200, Dec. 2001. doi: 10.1006/jsbi.2002.4435.

L. Rulísek and U. Ryde. Structure of reduced and oxidized manganese superoxide dismutase: a combined computational and experimental approach. *J. Phys. Chem. B*, 110 (23):11511–11518, June 2006. doi: 10.1021/jp057295t.

B. Rupp. *Biomolecular crystallography: Principles, Practice, and Application to Structural Biology*. CRC Press, Boca Raton, FL, Oct. 2009. ISBN 9780815340812.

U. Ryde. The coordination of the catalytic zinc ion in alcohol dehydrogenase studied by combined quantum-chemical and molecular mechanics calculations. *J. Comput. Aided Mol. Des.*, 10(2):153–164, Apr. 1996. doi: 10.1007/BF00402823.

U. Ryde. QM/MM calculations on proteins. *Methods in Enzymology*, 577:119–158, 2016. doi: 10.1016/bs.mie.2016.05.014.

U. Ryde and K. Nilsson. Quantum chemistry can locally improve protein crystal structures. *J. Am. Chem. Soc.*, 125(47):14232–14233, Nov. 2003. doi: 10.1021/ja0365328.

U. Ryde and M. H. M. Olsson. Structure, strain, and reorganization energy of blue copper models in the protein. *Int. J. Quantum Chem.*, 81(5):335–347, 2001. doi: 10.1002/1097-461X(2001)81:5<335::AID-QUA1003>3.0.CO;2-Q.

U. Ryde, L. Olsen, and K. Nilsson. Quantum chemical geometry optimizations in proteins using crystallographic raw data. *J. Comput. Chem.*, 23(11):1058–1070, Aug. 2002. doi: 10.1002/jcc.10093.

U. Ryde, Y.-W. Hsiao, L. Rulísek, and E. I. Solomon. Identification of the peroxy adduct in multicopper oxidases by a combination of computational chemistry and extended x-ray absorption fine-structure measurements. *J. Am. Chem. Soc.*, 129(4):726–727, Jan. 2007. doi: 10.1021/ja062954g.

A. Saha, S. S. Nia, and J. A. Rodríguez. Electron diffraction of 3D molecular crystals. *Chem. Rev.*, 122(17):13883–13914, Sept. 2022. doi: 10.1021/acs.chemrev.1c00879.

M. Sahlin, A. Gräslund, L. Petersson, A. Ehrenberg, and B. M. Sjöberg. Reduced forms of the iron-containing small subunit of ribonucleotide reductase from escherichia coli. *Biochemistry*, 28(6):2618–2625, Mar. 1989. doi: 10.1021/bi00432a039.

H. R. Saibil. Cryo-EM in molecular and cellular biology. *Mol. Cell*, 82(2):274–284, Jan. 2022. doi: 10.1016/j.molcel.2021.12.016.

R. C. Scarrow, M. J. Maroney, S. M. Palmer, L. Que, S. P. Salowe, and J. Stubbe. EXAFS studies of the B2 subunit of the ribonucleotide reductase from e. coli. *J. Am. Chem. Soc.*, 108(21):6832–6834, Oct. 1986. doi: 10.1021/ja00281a077.

S. H. W. Scheres. RELION: implementation of a bayesian approach to cryo-EM structure determination. *J. Struct. Biol.*, 180(3):519–530, Dec. 2012. doi: 10.1016/j.jsb.2012.09.006.

F. V. Schmidt, L. Schulz, J. Zarzycki, S. Prinz, N. N. Oehlmann, T. J. Erb, and J. G. Rebelein. Structural insights into the iron nitrogenase complex. *Nat. Struct. Mol. Biol.*, 31(1):150–158, Jan. 2024. doi: 10.1038/s41594-023-01124-2.

G. C. Schröder and F. Meilleur. Metalloprotein catalysis: structural and mechanistic insights into oxidoreductases from neutron protein crystallography. *Acta Crystallogr. D Struct. Biol.*, 77(Pt 10):1251–1269, Oct. 2021. doi: 10.1107/S2059798321009025.

Schrödinger. The PyMOL molecular graphics system, version 3.1.0, open-source build. https://github.com/schrodinger/pymol-open-source, 2024.

E. Schrödinger. Quantisierung als eigenwertproblem. *Ann. Phys.*, 385(13):437–490, Jan. 1926a. doi: 10.1002/andp.19263851302.

E. Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.*, 28(6):1049–1070, Dec. 1926b. doi: 10.1103/PhysRev.28.1049.

L. C. Seefeldt, Z.-Y. Yang, D. A. Lukoyanov, D. F. Harris, D. R. Dean, S. Raugei, and B. M. Hoffman. Reduction of substrates by nitrogenases. *Chem. Rev.*, 120(12):5082–5106, June 2020. doi: 10.1021/acs.chemrev.9b00556.

J. M. Seminario. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.*, 60(7):1271–1277, 1996. doi: 10.1002/(SICI)1097-461X(1996)60:7<1271::AID-QUA8>3.0.CO;2-.

T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, and A. C. T. van Duin. The ReaxFF reactive force-field: development, applications and future directions. *npj Comput. Mater.*, 2(1), Mar. 2016. doi: 10.1038/npjcompumats.2015.11.

H. M. Senn and W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed Engl.*, 48(7):1198–1229, 2009. doi: 10.1002/anie.200802019.

Y. Sheng, I. A. Abreu, D. E. Cabelli, M. J. Maroney, A.-F. Miller, M. Teixeira, and J. S. Valentine. Superoxide dismutases and superoxide reductases. *Chem. Rev.*, 114(7):3854–3918, Apr. 2014. doi: 10.1021/cr4005296.

F. Shu, V. Ramakrishnan, and B. P. Schoenborn. Enhanced visibility of hydrogen atoms by neutron crystallography on fully deuterated myoglobin. *Proc. Natl. Acad. Sci. U. S. A.*, 97(8):3872–3877, Apr. 2000. doi: 10.1073/pnas.060024697.

F. J. Sigworth. Principles of cryo-EM single-particle image processing. *Microscopy (Oxf.)*, 65(1):57–67, Feb. 2016. doi: 10.1093/jmicro/dfv370.

A. Singer. Mathematics for cryo-electron microscopy. 2018. doi: 10.48550/arXiv.1803.06714.

D. Sippel and O. Einsle. The structure of vanadium nitrogenase reveals an unusual bridging ligand. *Nat. Chem. Biol.*, 13(9):956–960, Sept. 2017. doi: 10.1038/nchembio.2428.

S. Sirajuddin, D. Barupala, S. Helling, K. Marcus, T. L. Stemmler, and A. C. Rosenzweig. Effects of zinc on particulate methane monooxygenase activity and structure. *J. Biol. Chem.*, 289(31):21782–21794, Aug. 2014. doi: 10.1074/jbc.M114.581363.

D. S. Sivia. *Data analysis*. Oxford University Press, June 2006. ISBN 9780198568322.

B. M. Sjöberg, T. M. Loehr, and J. Sanders-Loehr. Raman spectral evidence for a mu-oxo bridge in the binuclear iron center of ribonucleotide reductase. *Biochemistry*, 21(1): 96–102, Jan. 1982. doi: 10.1021/bi00530a017.

J. C. Slater. The self consistent field and the structure of atoms. *Phys. Rev.*, 32(3):339–348, Sept. 1928. doi: 10.1103/PhysRev.32.339.

J. C. Slater. The theory of complex spectra. *Phys. Rev.*, 34(10):1293–1322, Nov. 1929. doi: 10.1103/PhysRev.34.1293.

J. C. Slater. Note on hartree's method. *Phys. Rev.*, 35(2):210–211, Jan. 1930a. doi: 10.1103/PhysRev.35.210.2.

J. C. Slater. Atomic shielding constants. *Phys. Rev.*, 36(1):57–64, July 1930b. doi: 10.1103/PhysRev.36.57.

J. C. Slater. A simplification of the Hartree-Fock method. *Phys. Rev.*, 81(3):385–390, Feb. 1951. doi: 10.1103/PhysRev.81.385.

V. Smil. *Enriching the earth*. The MIT Press. MIT Press, London, England, Feb. 2004. ISBN 9780262693134.

S. M. Smith, S. Rawat, J. Telser, B. M. Hoffman, T. L. Stemmler, and A. C. Rosenzweig. Crystal structure and characterization of particulate methane monooxygenase from methylocystis species strain M. *Biochemistry*, 50(47):10231–10240, Nov. 2011. doi: 10.1021/bi200801z.

P. Söderhjelm and U. Ryde. Combined computational and crystallographic study of the oxidised states of [NiFe] hydrogenase. *J. Mol. Struct.*, 770(1):199–219, Sept. 2006. doi: 10.1016/j.theochem.2006.06.008.

M. Srnec, F. Aquilante, U. Ryde, and L. Rulísek. Reaction mechanism of manganese superoxide dismutase studied by combined quantum and molecular mechanical calculations and multiconfigurational methods. *J. Phys. Chem. B*, 113(17):6074–6086, Apr. 2009. doi: 10.1021/jp810247u.

P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.*, 98(45):11623–11627, Nov. 1994. doi: 10.1021/j100096a001.

J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria. Climbing the density functional ladder: nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, 91(14):146401, Oct. 2003. doi: 10.1103/PhysRevLett.91.146401.

G. Taylor. The phase problem. *Acta Crystallogr. D Struct. Biol.*, 59(11):1881–1890, Nov. 2003. doi: 10.1107/S0907444903017815.

G. L. Taylor. Introduction to phasing. *Acta Crystallogr. D Biol. Crystallogr.*, 66(Pt 4):325–338, Apr. 2010. doi: 10.1107/S0907444910006694.

T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, and P. D. Adams. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods*, 21(1):110–116, Jan. 2024. doi: 10.1038/s41592-023-02087-4.

R. F. Thompson, M. Walker, C. A. Siebert, S. P. Muench, and N. A. Ranson. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods*, 100:3–15, May 2016. doi: 10.1016/j.ymeth.2016.02.017.

R. Tian, Y. Geng, H. Guo, C. Yang, I. Seim, and G. Yang. Comparative analysis of the superoxide dismutase gene family in cetartiodactyla. *J. Evol. Biol.*, 34(7):1046–1060, July 2021. doi: 10.1111/jeb.13792.

I. J. Tickle. Statistical quality indicators for electron-density maps. *Acta Crystallogr. D Struct. Biol.*, 68(Pt 4):454–467, Apr. 2012. doi: 10.1107/S0907444911035918.

C. Trncik, F. Detemple, and O. Einsle. Iron-only fe-nitrogenase underscores common catalytic principles in biological nitrogen fixation. *Nat. Catal.*, 6(5):415–424, Apr. 2023. doi: 10.1038/s41929-023-00952-1.

D. E. Tronrud. TNT refinement package. *Methods in Enzymology*, 277:306–319, 1997. ISSN 0076-6879. doi: 10.1016/S0076-6879(97)77017-4.

D. E. Tronrud. Introduction to macromolecular refinement. *Acta Crystallogr. D Struct. Biol.*, 60(Pt 12 Pt 1):2156–2168, Dec. 2004. doi: 10.1107/S090744490402356X.

K. N. Trueblood, H. B. Bürgi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schulz, U. Shmueli, and S. C. Abrahams. Atomic dispacement parameter nomenclature. report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr. A*, 52(5):770–781, Sept. 1996. doi: 10.1107/S0108767396005697.

O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller. Machine learning force fields. *Chem. Rev.*, 121(16): 10142–10186, Aug. 2021. doi: 10.1021/acs.chemrev.0c01111.

A. Urzhumtsev and V. Y. Lunin. Analytic modeling of inhomogeneous-resolution maps in cryo-electron microscopy and crystallography. *IUCrJ*, 9(Pt 6):728–734, Nov. 2022. doi: 10.1107/S2052252522008260.

A. Urzhumtsev, P. V. Afonine, V. Y. Lunin, T. C. Terwilliger, and P. D. Adams. Metrics for comparison of crystallographic maps. *Acta Crystallogr. D Struct. Biol.*, 70(Pt 10): 2593–2606, Oct. 2014. doi: 10.1107/S1399004714016289.

A. G. Urzhumtsev and V. Y. Lunin. Introduction to crystallographic refinement of macromolecular atomic models. *Crystallogr. Rev.*, 25(3):164–262, July 2019. doi: 10.1080/0889311X.2019.1631817.

A. G. Urzhumtsev, T. P. Skovoroda, and V. Y. Lunin. A procedure compatible with *X-PLOR* for the calculation of electron-density maps weighted using an *R*-free-likelihood approach. *J. Appl. Crystallogr.*, 29(6):741–744, Dec. 1996. doi: 10.1107/S0021889896007194.

A. G. Urzhumtsev, L. M. Urzhumtseva, and V. Y. Lunin. Direct calculation of cryo-EM and crystallographic model maps for real-space refinement. *Acta Crystallogr. D Struct. Biol.*, 78(Pt 12):1451–1468, Dec. 2022. doi: 10.1107/s2059798322010907.

L. Urzhumtseva, V. Lunin, and A. Urzhumtsev. Algorithms and programs for the shell decomposition of oscillating functions in space. *J. Appl. Crystallogr.*, 56(1):302–311, Feb. 2023. doi: 10.1107/S160057672201144X.

L. Valenti, D. Conte, A. Piperno, P. Dongiovanni, A. L. Fracanzani, M. Fraquelli, A. Vergani, C. Gianni, L. Carmagnola, and S. Fargion. The mitochondrial superoxide dismutase A16V polymorphism in the cardiomyopathy associated with hereditary haemochromatosis. *J. Med. Genet.*, 41(12):946–950, Dec. 2004. doi: 10.1136/jmg.2004.019588.

M. Van Heel. Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21(2):111–123, Jan. 1987. doi: 10.1016/0304-3991(87)90078-7.

F. S. Varley. Neutron scattering lengths and cross sections. *Neutron News*, 3(3):26–37, Jan. 1992. doi: 10.1080/10448639208218770.

J. L. Vilas, J. M. Carazo, and C. O. S. Sorzano. Emerging themes in CryoEM⊠Single particle analysis image processing. *Chem. Rev.*, 122(17):13915–13951, Sept. 2022. doi: 10.1021/acs.chemrev.1c00850.

J. G. Vilhena, L. Greff da Silveira, P. R. Livotto, I. Cacelli, and G. Prampolini. Automated parameterization of quantum mechanically derived force fields for soft materials and complex fluids: Development and validation. *J. Chem. Theory Comput.*, 17(7):4449–4464, July 2021. doi: 10.1021/acs.jctc.1c00213.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, Mar. 2020. doi: 10.1038/s41592-019-0686-2.

S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, Aug. 1980. doi: 10.1139/p80-159.

V. P. Vysotskiy, M. Torbjörnsson, H. Jiang, E. D. Larsson, L. Cao, U. Ryde, H. Zhai, S. Lee, and G. K.-L. Chan. Assessment of DFT functionals for a minimal nitrogenase [Fe(SH)4H]- model employing state-of-the-art ab initio methods. *J. Chem. Phys.*, 159 (4), July 2023. doi: 10.1063/5.0152611.

I. Waller. Zur frage der einwirkung der wärmebewegung auf die interferenz von röntgenstrahlen. *Eur. Phys. J. A*, 17(1):398–408, Dec. 1923. doi: 10.1007/BF01328696.

J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, July 2004. doi: 10.1002/jcc.20035.

J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, Oct. 2006. doi: 10.1016/j.jmgm.2005.12.005.

L. Wang, H. Kruse, O. V. Sobolev, N. W. Moriarty, M. P. Waller, P. V. Afonine, and M. Biczysko. Real-space quantum-based refinement for cryo-EM: Q|R#3. *bioRxivorg*, May 2020. doi: 10.1101/2020.05.25.115386.

V. C.-C. Wang, S. Maji, P. P.-Y. Chen, H. K. Lee, S. S.-F. Yu, and S. I. Chan. Alkane oxidation: Methane monooxygenases, related enzymes, and their biomimetics. *Chem. Rev.*, 117(13):8574–8621, July 2017. doi: 10.1021/acs.chemrev.6b00624.

Y. Wang, H. Kruse, N. W. Moriarty, M. P. Waller, P. V. Afonine, and M. Biczysko. Optimal clustering for quantum refinement of biomolecular structures: Q|R#4. *Theor. Chem. Acc.*, 142(10), Oct. 2023. doi: 10.1007/s00214-023-03046-0.

A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, May 1976. doi: 10.1016/0022-2836(76)90311-9.

C. X. Weichenberger, P. V. Afonine, K. Kantardjieff, and B. Rupp. The solvent component of macromolecular crystals. *Acta Crystallogr. D Struct. Biol.*, 71(Pt 5):1023–1038, May 2015. doi: 10.1107/S1399004715006045.

F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, Sept. 2005. doi: 10.1039/B508541A.

M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1):160018, Mar. 2016. doi: 10.1038/sdata.2016.18.

M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Struct. Biol.*, 67(Pt 4):235–242, Apr. 2011. doi: 10.1107/S0907444910045749.

M. Wojdyr. GEMMI: A library for structural biology. *J. Open Source Softw.*, 7(73):4200, May 2022. doi: 10.21105/joss.04200.

wwPDB Consortium. Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, 47(D1):D520–D528, Jan. 2019. doi: 10.1093/nar/gky949.

wwPDB Consortium. EMDB-the electron microscopy data bank. *Nucleic Acids Res.*, 52 (D1):D456–D465, Jan. 2024. doi: 10.1093/nar/gkad1019.

wwPDB Consortium. EMDB-the electron microscopy data bank. https://www.ebi.ac.uk/emdb, 2025a. [Online; accessed 2025-09-02].

wwPDB Consortium. PDB-the protein data bank. https://www.rcsb.org, 2025b. [Online; accessed 2025-09-01].

Z. Yan, X. Li, and L. W. Chung. Multiscale quantum refinement approaches for metalloproteins. *J. Chem. Theory Comput.*, 17(6):3783–3796, June 2021. doi: 10.1021/acs.jctc.1c00148.

Z. Yan, D. Wei, X. Li, and L. W. Chung. Accelerating reliable multiscale quantum refinement of protein-drug systems enabled by machine learning. *Nat. Commun.*, 15(1):4181, May 2024. doi: 10.1038/s41467-024-48453-4.

Z. Yang, X. Zeng, Y. Zhao, and R. Chen. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target. Ther.*, 8(1):115, Mar. 2023. doi: 10.1038/s41392-023-01381-z.

N. Yu, H. P. Yennawar, and K. M. Merz, Jr. Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. *Acta Crystallogr. D Struct. Biol.*, 61(Pt 3):322–332, Mar. 2005. doi: 10.1107/S0907444904033669.

X. Zhang, B. B. Ward, and D. M. Sigman. Global nitrogen cycle: Critical enzymes, organisms, and processes for nitrogen budgets and dynamics. *Chem. Rev.*, 120(12):5308–5351, June 2020. doi: 10.1021/acs.chemrev.9b00613.

Y. Zhao, S.-M. Bian, H.-N. Zhou, and J.-F. Huang. Diversity of nitrogenase systems in diazotrophs. *J. Integr. Plant Biol.*, 48(7):745–755, July 2006. doi: 10.1111/j.1744-7909.2006.00271.x.

H. Zheng, D. R. Cooper, P. J. Porebski, I. G. Shabalin, K. B. Handing, and W. Minor. CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr. D Struct. Biol.*, 73(Pt 3):223–233, Mar. 2017a. doi: 10.1107/S2059798317001061.

M. Zheng, N. W. Moriarty, Y. Xu, J. R. Reimers, P. V. Afonine, and M. P. Waller. Solving the scalability issue in quantum-based refinement: Q|R#1. *Acta Crystallogr. D Struct. Biol.*, 73(Pt 12):1020–1028, Dec. 2017b. doi: 10.1107/S2059798317016746.

M. Zheng, J. R. Reimers, M. P. Waller, and P. V. Afonine. Q|R: quantum-based refinement. *Acta Crystallogr. D Struct. Biol.*, 73(Pt 1):45–52, Jan. 2017c. doi: 10.1107/S2059798316019847.

M. Zheng, M. Biczysko, Y. Xu, N. W. Moriarty, H. Kruse, A. Urzhumtsev, M. P. Waller, and P. V. Afonine. Including crystallographic symmetry in quantum-based refinement: Q|R#2. *Acta Crystallogr. D Struct. Biol.*, 76(Pt 1):41–50, Jan. 2020. doi: 10.1107/S2059798319015122.

M. Zheng, Y. Liu, G. Zhang, Z. Yang, W. Xu, and Q. Chen. The applications and mechanisms of superoxide dismutase in medicine, food, and cosmetics. *Antioxidants (Basel)*, 12(9), Aug. 2023. doi: 10.3390/antiox12091675.

R. Zubatyuk, M. Biczysko, K. Ranasinghe, N. W. Moriarty, H. Gokcan, H. Kruse, B. K. Poon, P. D. Adams, M. P. Waller, A. E. Roitberg, O. Isayev, and P. V. Afonine. AQuaRef: machine learning accelerated quantum refinement of protein structures. *Nat. Commun.*, 16(1):9224, Oct. 2025. doi: 10.1038/s41467-025-64313-1.

# Scientific publications

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

ISBN: 978-91-8096-130-1

Division of Computational Chemistry
Department of Chemistry
Faculty of Science
Lund University